Doctorat en Societat de la Informació
i el Coneixement
Universitat Oberta de Catalunya (UOC)

TESI DOCTORAL

AVALUACIÓ EX-ANTE I EX-POST DE LA RECERCA: APROXIMACIONS PER A ANALITZAR LA VALIDESA DE DUES EINES D'AVALUACIÓ DE LA RECERCA

Maite Solans Domènech

Directora: Dra. Marta Aymerich

Juliol 2020

What gets measured gets improved

Peter F. Drucker

Resum

L'objectiu general d'aquesta tesi ha estat desenvolupar diverses estratègies per validar dos mètodes d'avaluació de la recerca que es porten a terme en dues fases del procés de recerca.

El primer estudi està relacionat amb l'avaluació ex-ante, que implica l'avaluació prèvia a l'atorgament d'ajuts, en aquest cas, de projectes de recerca en ciències de la salut. El procés d'avaluació per parells analitzat consta de dues fases. En una primera fase s'avalua el projecte anonimitzat (sense identificar investigadors ni les seves institucions), mentre que en la segona fase, la identitat dels investigadors i de les seves institucions es revela. S'ha avaluat el canvi entre la primera valoració (projecte anonimitzat) i la segona valoració (no anonimitzat) per a cada avaluació i avaluador. També s'ha analitzat els factors relacionats amb el canvi, tenint en compte les característiques dels projectes, dels avaluadors, i dels investigadors. Per últim s'ha realitzat una anàlisi qualitativa del contingut dels comentaris dels avaluadors per conèixer els motius del canvi. L'anàlisi de 5.002 avaluacions va indicar que en quasi el 19% de les avaluacions, l'avaluador canviava la valoració, bé per millorar-la (12%) o per empitjorar-la (7%). Les troballes també suggereixen que aquest canvi està principalment correlacionat amb una avaluació positiva o negativa de l'experiència de l'equip investigador principal, a més d'altres factors. En resum, anonimitzar la identitat dels investigadors i les seves institucions en una primera fase del procés d'avaluació, pot ajudar a centrar-se més en el projecte i reduir alguns dels biaixos comuns del procés de revisió per parells.

El segon estudi està relacionat amb una avaluació ex-post en l'àmbit universitari, que implica l'avaluació de projectes de recerca un cop finalitzats, en aquest cas, amb l'avaluació del seu impacte en la societat. S'ha volgut avaluar la validesa d'un qüestionari per mesurar els impactes percebuts d'una àmplia gamma de projectes de recerca (arts i humanitats, ciències socials, ciències de la salut i tecnologies de la informació i la comunicació). També s'ha avaluat els impactes i les característiques associades. El qüestionari demostra una bona consistència interna i unes valideses de contingut i discriminants acceptables. No obstant això, les seves propietats mètriques són més poderoses en àmbits on tradicionalment s'ha avaluat la recerca (resums a congressos, articles científics o formació d'investigadors), àmbits on els investigadors tenen un cert control i influència. En general, els projectes de recerca van representar un estímul per a la producció de coneixement i el desenvolupament de competències de recerca. Els

aspectes conductuals, com ara la participació dels usuaris potencials o els projectes orientats a la missió (orientats a aplicacions pràctiques) es van associar amb majors beneficis socials.

L'anàlisi de la validesa de diferents enfocaments d'avaluació de la recerca pot contribuir de manera important a la credibilitat i acceptabilitat de les eines d'avaluació utilitzades. Els diferents enfocaments d'avaluació de la recerca poden proporcionar una informació molt valuosa per a la presa de decisions, sempre que es tinguin en compte alguns reptes que poden limitar la seva validesa i, per tant, la seva credibilitat i que inclourien: la importància de l'experiència dels experts utilitzats; la contextualització del procés d'avaluació; els interessos dels grups que han d'utilitzar els procediments o els seus resultats; definicions clares dels conceptes utilitzats; identificar quins elements es poden percebre com a facilitadors o barreres per a un resultat determinat; i sense oblidar, confirmar la qualitat de les dades.

En conclusió, analitzar la validesa dels mètodes d'avaluació de la recerca és una mirada cap endavant que ajuda a entendre i millora el procés d'avaluació i la seva credibilitat i, per tant, la utilitat de les mateixes avaluacions.

Resumen

El objetivo general de esta tesis ha sido desarrollar diversas estrategias para validar dos métodos de evaluación de la investigación que se llevan a cabo en dos fases del proceso de investigación.

El primer estudio está relacionado con la evaluación ex-ante, que implica la evaluación previa a la concesión de ayudas, en este caso, de proyectos de investigación en ciencias de la salud. El proceso de evaluación por pares analizado consta de dos fases. En una primera fase se evalúa el proyecto anonimizado (sin identificar investigadores ni sus instituciones), mientras que en la segunda fase, la identidad de los investigadores y de sus instituciones se revela. Se ha evaluado el cambio entre la primera valoración (proyecto anonimizado) y la segunda valoración (no anonimizado) para cada evaluación y evaluador. También se ha analizado los factores relacionados con el cambio, teniendo en cuenta las características de los proyectos, los evaluadores, y de los investigadores. Por último se ha realizado un análisis cualitativo del contenido de los comentarios de los evaluadores para conocer los motivos del cambio. El análisis de 5.002 evaluaciones indicó que en casi el 19% de las evaluaciones, el evaluador cambiaba la valoración, bien para mejorarla (12%) o para empeorarla (7%). Los hallazgos también sugieren que este cambio está principalmente correlacionado con una evaluación positiva o negativa de la experiencia del equipo investigador principal, aparte de otros factores. En resumen, anonimizar la identidad de los investigadores y sus instituciones en una primera fase del proceso de evaluación, puede ayudar a centrarse más en el proyecto y reducir algunos de los sesgos comunes del proceso de revisión por pares.

El segundo estudio está relacionado con una evaluación ex-post en el ámbito universitario, que implica la evaluación de proyectos de investigación una vez finalizados, en este caso, con la evaluación del impacto social. Se ha querido evaluar la validez de un cuestionario para medir los impactos percibidos de una amplia gama de proyectos de investigación (artes y humanidades, ciencias sociales, ciencias de la salud y tecnologías de la información y la comunicación). También se ha evaluado los impactos y las características asociadas. El cuestionario demuestra una buena consistencia interna y una validez de contenido y discriminante aceptables. Sin embargo, sus propiedades métricas son más poderosas en ámbitos donde tradicionalmente se ha evaluado la ciencia (resúmenes a congresos, artículos científicos o formación de investigadores), ámbitos donde los investigadores tienen un cierto control e influencia. En general, los proyectos de investigación representaron un estímulo para la producción de

conocimiento y el desarrollo de competencias de investigación. Los aspectos conductuales, como la participación de los usuarios potenciales o los proyectos orientados a la misión (orientados a aplicaciones prácticas) se asociaron con mayores beneficios sociales.

El análisis de la validez de diferentes enfoques de evaluación de la investigación puede contribuir de manera importante a la credibilidad y aceptabilidad de las herramientas de evaluación utilizadas. Los diferentes enfoques de evaluación de la investigación pueden proporcionar una información muy valiosa para la toma de decisiones, siempre que se tengan en cuenta algunos retos que pueden limitar su validez y, por tanto, su credibilidad: la importancia de la experiencia de los expertos utilizados; la contextualización del proceso de evaluación; los intereses de los grupos que han de utilizar los procedimientos o sus resultados; definiciones claras de los conceptos utilizados; identificar qué elementos se pueden percibir como facilitadores o barreras para un resultado determinado; y sin olvidar, confirmar la calidad de los datos.

En conclusión, analizar la validez de los métodos de evaluación de la investigación es una mirada hacia adelante que ayuda a entender y mejora el proceso de evaluación y su credibilidad y, por tanto, la utilidad de las mismas evaluaciones.

English abstract

The overall aim of this thesis was to validate two research evaluation methodologies that appear in two diferent phases of the research process.

The first study was related to the ex-ante evaluation, which involves the evaluation procedure for the selection of the proposals to be funded, in our case, proposals related to health sciences research. The study validated a two-phase peer review process. In the first phase, a blinded assessment, where researchers linked to a research proposal and the institutions they represent are concealed, while in the second assessment, the identity of the researchers and their institutions is revealed. The change made between the first (researchers/institutions blinded) and the second assessments (unblinded) was assessed for each evaluation and for each reviewer. Factors related to change were also analyzed, taking into account the characteristics of projects, reviewers, and researchers. Finally, a qualitative analysis of the content of the reviewers' comments was carried out to find out the reasons for the change. The analysis of 5,002 evaluations indicated that in 19% of the evaluations, the reviewer changed the second assessment: either for better (12%) or worse (7%). Our findings also suggest that a change in the second assessment was highly correlated with a positive evaluation of the experience of the principal investigator or research team, although other factors were also correlated. So, blinding the identity of researchers and their institutions in an early stage of the evaluation process can help to focus exclusively on the proposal and reduce some of the common biases of the peer-review process in grant decisions.

The second study is related to an ex-post evaluation, which involves the evaluation of research projects after completion. The aim was to evaluate the validity of an instrument designed to measure the perceived impacts of a wide range of research projects (arts and humanities, social sciences, health sciences and information and communication technologies) at a university level. The impacts perceived and their associated characteristics were also assessed. The easy-to-use questionnaire developed demonstrated good internal consistency and acceptable content validity. However, its metric properties were more powerful in areas where research is traditionally been evaluated (such as conference abstracts, scientific papers, or researcher training), and in areas where researchers in which the researchers had a degree of control and influence. In general, the research projects represented an stimulus for the production of knowledge and the development of research skills. Behavioural aspects such as engagement

with potential users or mission-oriented projects (targeted to practical applications) were associated with higher social benefits.

Analyzing the validity of different research evaluation approaches (ex ante and ex post) can make an important contribution to the credibility and acceptability of the evaluation tools used. Different approaches to research evaluation can provide very valuable information for decision-making. Some challenges should be taken into account that may limit its validity and therefore its credibility such as the experience and competence of the experts used in the evaluation process; the contextualization in which the evaluation process takes place; the needs of the stakeholders interested in the procedure or their outcomes; a clear definition of the concepts used in the evaluation; identify barriers and facilitators of a given outcome; and finally, ensure tdata quality.

In conclusion, analyzing the validity of research evaluation procedures is a forward-looking approach that helps to understand and improve the evaluation process and its credibility, and therefore the usefulness of the evaluations.

Agraïments

Tot i que una tesi és el resultat d'un treball individual i de l'experiència personal, aquest no haurà estat possible sense els consells i aportacions d'altres persones que, sens dubte, han ajudat a millorar el resultat final. A totes elles, el meu més sincer agraïment i acompanyament en aquest camí.

En primer lloc, agrair a la meva directora, la doctora Marta Aymerich, la seva orientació, comprensió, paciéncia i sobretot ànims. En segon lloc, als meus companys de feina, Paula Adam i Joan MV Pons, per ensenyar-me tant dia a dia. A la meva companya i amiga, Mercé Obach, per la seva aportació i comentaris útils. També estic molt agraït als coautors de les publicacions d'aquesta tesi; Carme Carrion, Ignacio Ferreira-Gonzáalez, Josep Grau, Imma Guillamón, Gaietà Permanyer-Miralda i Aida Ribera, per la nostra fructífera col·laboració.

I per últim, no puc obviar la família i les amistats, pels seus suports i ànims interminables.

Abreviatures

D: Dubtós

F: Finançable

FR: Finançable amb Reserves

IC: Interval de Confiança

ID: Identificador del projecte

IP: Investigador Principal

NF = No Finançable

RRR: Reducció Relativa del Risc

TIC: Tecnologies de la Informació i la Comunicació

UOC: Universitat Oberta de Catalunnya

Table of Contents

Capítol 1: Antecedents i Objectius

JUSTIFICACIÓ

L'avaluació de la qualitat de la recerca no és cap novetat. Els investigadors han debatut durant molt de temps els millors criteris i mitjans per determinar el rigor científic i la importància de la recerca.[1] El que és nou és la preocupació creixent en les institucions acadèmiques, els governs, les agències de finançament de la recerca i les entitats sense ànim de lucre quant a la rellevància de la recerca que es finança, en el sentit d'usabilitat i d'impacte. A més, l'interès i la demanda d'avaluacions de la recerca són pràctiques cada vegada més arrelades en l'àmbit internacional i vinculades cada vegada més a demostrar la responsabilitat i la transparència dels seus processos.[2] La disminució de la inversió en recerca i l'augment de la competència per aconseguir fons fa que això sigui encara més crític. A més, també existeix una preocupació creixent de com augmentar el valor de la recerca i com reduir recerca ineficaç, aquella redundant o que no aporta res de valor.[3] Totes aquestes preocupacions són claus per determinar que el finançament es destina a científics excel·lents que produeixen resultats excel·lents i socialment valuosos.[4] De fet, la mateixa definició de recerca, que segons el manual de Frascati, ja inclou tots aquests aspectes en el seu redactat: la recerca és 'un treball creatiu realitzat de manera sistemàtica per augmentar l'estoc de coneixement, inclòs el coneixement de l'home, la cultura i la societat, i l'ús d'aquest inventari de coneixements per idear noves aplicacions [5]

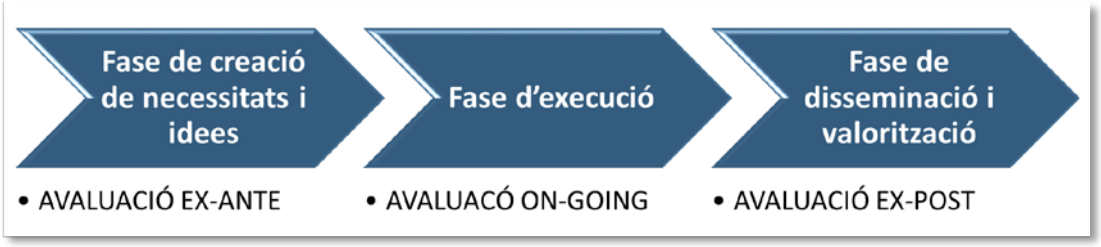
Internacionalment hi ha països pioners com Gran Bretanya, Canadà, els Països Baixos o Austràlia on el paper de l'avaluació com a agent transformador és clau. En aquests països, la política científica, igual que altres polítiques públiques, cada vegada es basa més en l'evidència científica que ofereixen els processos avaluadors. Per tant, ens trobem en un entorn creixent de processos orientats a millorar l'efectivitat de la recerca mitjançant l'avaluació de les polítiques i les pràctiques científiques (recerca de la recerca). Això ha donat lloc a un gran nombre de processos, instruments i indicadors la qual cosa comporta un problema creixent entorn de la validesa d'aquests processos. S'ha d'analitzar què funciona i què no funciona, i explicar i compartir casos d'èxit de manera transparent, així com promoure que l'entorn, el sistema i els destinataris dels resultats científics en siguin coneixedors i beneficiaris.

ESTAT DE LA QÜESTIÓ

Concepte d'avaluació de la recerca

El concepte d'avaluació de la recerca es pot definir com una avaluació sistemàtica de l'activitat, el procés o el resultat de recerca en totes les seves variants.[6] Que sigui 'sistemàtica' implica que hi ha una elevada exigència sobre els mètodes utilitzats, incloent-hi l'ús de metodologies apropiades i d'última generació i processos sòlids que siguin vàlids. La Comissió Europea defineix l'avaluació com 'un judici crític basat en evidències'.[7] A la pràctica, l'avaluació de la recerca es pot aplicar a molts nivells i en diferents fases del procés de recerca (Gràfic 1), pot prendre moltes formes i utilitzar diversos mètodes.

Gràfic 1. Avaluació de la recerca en diferents moments del procés



En primer lloc, l'avaluació de la recerca es pot realitzar en la fase de planificació d'un programa o projecte (avaluació ex-ante). Es tracta de decisions estratègiques que inclouen d'una banda, l'avaluació de necessitats que determina el disseny de programes o projectes de recerca i de l'altra, l'avaluació associada a l'atorgament de recursos. L'avaluació de necessitats de recerca incorpora eines d'avaluació utilitzades tant per a detectar necessitats temàtiques (o àrees de coneixement o reptes socioeconòmics) com tipologies d'accions (persones, idees, infraestructures ...) a finançar. L'avaluació associada a l'atorgament de recursos es fa majoritàriament per via de règim de concurrència competitiva, i basant-se en uns criteris i metodologia preestablerts que constitueixen el ben conegut procés de revisió per consemblants (*peer-review*). Aquesta avaluació ex-ante està relacionada amb la coherència i la rellevància de la recerca, la solidesa metodològica i el realisme dels efectes esperats. En l'avaluació associada a l'atorgament de recursos, les avaluacions per consemblants són el sistema més utilitzat. Malgrat el seu ús generalitzat, les deficiències en el procés per revisió per consemblants, en termes d'eficiència i eficàcia, estan ben documentats.[8,9] Tot i ser imperfecte, és un instrument acceptat per a l'avaluació ex-ante de la qualitat de la recerca.[10]

El seguiment de les activitats de les propostes finançades (avaluació on-going) pot ser un procés continu destinat, no només a la supervisió, sinó com un exercici de suport i acompanyament per l'assoliment dels objectius individuals. L'avaluació on-going té l'objectiu d'orientar la recerca a mig camí, aborda les qüestions operatives, incloent-hi l'adopció de 'bones pràctiques'. Tot i que no deixa de ser rellevant, aquesta avaluació s'ha realitzat tradicionalment com un exercici administratiu, i per tant, no s'inclourà dins del treball d'aquesta tesi.

Finalment, l'avaluació de la recerca es pot realitzar quan els projectes d'un programa s'han completat (avaluació ex-post). Un exemple podria ser l'avaluació de l'impacte social dels resultats de la recerca (avaluació de resultats). L'objectiu de l'avaluació ex-post és principalment demostrar els diferents efectes de la recerca. L'avaluació de l'impacte de la recerca s'ha desenvolupat majoritàriament en països anglosaxons amb diferents exemples de publicacions científiques, mètodes i marcs que demostren la viabilitat d'emprendre aquestes anàlisis. [11–15] Les avaluacions s'han fet principalment en ciències de la salut, tot i que altres disciplines s'han anat afegint a aquestes avaluacions, sobretot en recerca agro-ambiental[16] i més recentment, en recerca en arts i humanitats.[17–19] A causa de la creixent demanda i aspiracions per avaluar l'impacte de la recerca (més enllà de l'avaluació tradicional basada en les publicacions acadèmiques), el 2018 es va publicar l'*ISRIA statement*,[20] un document on s'inclouen deu recomanacions per a poder realitzar un procés rigorós i efectiu d'avaluació de l'impacte de la recerca. Una de les principals recomanacions és que és molt important conèixer quin és l'objectiu de l'avaluació, podent-se definir diferents motius per utilitzar les avaluacions de la recerca: per decidir sobre l'assignació de recursos; per donar suport a processos interns d'anàlisi sobre el sistema de recerca; per demostrar que la realització de la recerca és responsable, relacionat també amb la transparència en la inversió de diners; i per demostrar el valor de la recerca i els seus processos. D'acord amb això, els resultats de l'avaluació són utilitzables per a la presa de decisions i d'aquesta manera la presa de decisions està basada en l'evidència (el que es coneix com a *evidence-based decision making*).[21]

Validesa dels processos i eines d'avaluació de la recerca

La validesa és un aspecte clau de tota avaluació. Una atenció meticulosa a aquest aspecte pot fer que s'acceptin millor els seus resultats i, per tant, la seva credibilitat i confiança. El concepte de validesa és polifacètic. D'una banda, les definicions més clàssiques el relacionen amb l'eina de mesura, el procés, o la riquesa i abast de les dades obtingudes. [22] En canvi, les definicions

més modernes de validesa suggereixen que la validació és un procés on es pretén validar els resultats de l'avaluació, i no pas l'eina o el procés.[23] La correspondència entre el constructe que es vol mesurar i el contingut de l'instrument (els diferents ítems de l'eina) és també una part important d'aquestes definicions més modernes de validesa. Inclou per tant, que l'instrument mesuri aspectes clau del constructe que es vol mesurar, i que els resultats de l'avaluació no estiguin afectats per altres factors, fora d'aquells pels quals l'instrument s'ha dissenyat.[24] Tanmateix resulta difícil que una avaluació sigui 100% vàlida, és a dir, difícilment s'esvaïran completament totes les amenaces que afecten la validesa. Per tant, l'objectiu ha d'estar més centrat a atenuar aquestes amenaces i minimitzar la 'invalidesa' més que maximitzar la validesa interna de l'avaluació.

Validesa del procés d'avaluació ex-ante de la recerca

La revisió per consemblants (*peer review*) és el procés més utilitzat en les avaluacions ex-ante de propostes per tal de pendre decisions sobre els finançament de la recerca.[25] Malgrat el seu extens ús, la validesa del seu procés, principalment quant al resultat obtingut i la seva credibilitat, sovint reben crítiques.[8] Aquestes, s'han anat accentuant a mesura que la taxa d'èxit de rebre finançament va disminuint.[26]

Dins de la validesa dels processos d'avaluació ex-ante, podem diferenciar quatre tipus de valideses, relacionades estretament amb les crítiques més accentuades al procés.[8]

Validesa en el resultat obtingut

Valora fins a quin punt, el resultat obtingut amb el procés d'avaluació aconsegueix el resultat desitjat. En funció dels objectius dels finançadors, es pot determinar tant per si el procés aconsegueix els resultats desitjats, la qualitat de les propostes, o la rellevància social de les mateixes.

Validesa en la flexibilitat del resultat

Valora si el procés és prou flexible i dóna suport a diferents tipus de recerca, per exemple, impulsant la innovació, el treball interdisciplinari, translacional i/o aplicat, o pel contrari, afavoreix el conservacionisme.

Validesa en la fiabilitat de la presa de decisions

Valora la credibilitat del procés, en el sentit de que en aquest no es determinin biaixos relacionats amb certes disciplines, organitzacions, tipus de propostes o característiques dels sol·licitants com el gènere. Si l'avaluació es fa per consemblants també inclou la concordança entre les avaluacions dels diferents revisors.

Validesa en la responsabilitat de la presa de decisions

Valora la transperència del procés, en el sentit de l'adeqüació i claretat en la manera com es prenen les decisions de finançament.

L'adequació de diferents enfocaments per millorar la validesa del procés d'avaluació ex-ante, depenen de la missió del finançador, del finançament disponible, o de la recerca que es vol finançar. Algunes opcions utilitzades per millorar la validesa del procés d'avauació inclouen la formació dels avaluadors, la normalització dels procediments, la disseminació oberta del procés, la rigorositat en la cerca i tria d'avaluadors, la utilització de coordinadors que revisen les tasques dels avaluadors, o l'anonimització dels projectes, entre d'altres. En aquesta tesi es treballaran dos dels conceptes: la validesa en el resultat obtingut i la validesa en la fiabilitat de la presa de decisions que han de permetre validar la qualitat del procés d'una revisió per consemblants. Validar el procés de revisió no és més que dividir-lo en petites fases i comprovar què passa en aquestes fases, per exemple, si incloem una fase on s'avaluen propostes anonimitzades.

L'anonimització dels projectes ha estat sovint un tema de debat en la validesa i adequació de les avaluacions. Tot i que la majoria de l'evidència disponible sobre l'efecte de l'anonimització en la qualitat de les avaluacions estan associades amb el sistema d'avaluació de les publicacions enviades al sistema editorial de revisió per consemblants,[27] alguns estudis s'han centrat en l'anonimització de les propostes de recerca, amb resultats no concordants.

Un dels estudis més coneguts de revisió per consemblants de propostes de recerca es va realitzar entre els anys 70 i 80 a la National Science Foundation (NSF), als EUA. La meitat dels revisors van rebre propostes que havien estat anonimitzades, mentre que l'altra meitat rebia còpies idèntiques que no s'havien anonimitzat. Els resultats de l'estudi de validació va demostrar que l'edat i l'historial científic del sol·licitant tenien poc efecte en les possibilitats d'obtenir una subvenció i que per tant no es trobaben biaixos en l'assignació de recursos.[28]

Tot i aquest resultat, des d'aleshores, més evidència s'ha anat creant indicant que l'anonimització dels investigadors i les seves institucions pot reduir biaixos davant de determinades característiques i ajuda a evitar qualsevol conflicte d'interès interpersonal.[29] D'una banda, característiques dels investigadors i les seves institucions; com el sexe o l'edat; però especialment el nom i renom dels investigadors i les institucions; o l'efecte Mateu, que implica donar més finançament a aquells que ja tenen finançament,[30,31] o l'efecte Matilda, un biax on l'èxit dels treballs científicsportats a terme per dones, s'acaba atribuint als seus col·legues científics homes.[32] D'altra, característiques de la proposta; molt sovint relacionades amb propostes innovadores o interdisciplinàries.[33] Per exemple, en un estudi més recent es va demostrar que, tot i controlar per l'excel·lència científica del sol·licitant, s'observava que característiques del sol·licitant, de les propostes i del revisor podien influir en l'avaluació ex-ante.[34] Entre d'altres, es van trobar pitjors puntuacions per propostes de ciència aplicada o desigualtats de gènere a favor dels homes. Els autors concluïen que la combinació de diferents característiques que portaven a biaix podien tenir un efecte substancial en la puntuació d'una proposta i convertir-la d'una proposta finançable a una no finançable.

Per tots aquests motius i seguint les recomanacions de la literatura,[26] s'ha de reconèixer, avaluar i analitzar la incertesa i la validesa dels processos d'avaluació per consemblants.

Validesa de les eines d'avaluació ex-post de la recerca

Els qüestionaris són un dels mètodes més utilitzats en l'avaluació ex-post de l'impacte de la recerca, ja que proporcionen una visió àmplia de l'estat d'un cos de recerca i proporcionen dades i informació comparables i fàcils d'analitzar. La normalització o estandardització de l'enfocament millora aquesta comparabilitat i minimitza el biaix i la subjectivitat de l'investigador, especialment en el cas d'enquestes web o postals. Una acurada construcció de les preguntes augmenta la validesa dels resultats.[2]

Per a mesures de resultat com els qüestionaris, la definició tradicional de validesa, diu que ser 'vàlid' significa, d'una banda, que les eines d'avaluació mesuren allò que es vol mesurar (validesa interna) i que ho fan amb precisió, és a dir, sense variació en repetir la mesura (fiabilitat).[22] És a dir, ens referim bàsicament a característiques de l'eina. Tot i que la bibliografia classifica la validesa de diferents maneres, podem dir que la validesa té tres grans components: el contingut, el criteri i el concepte:[38]

Validesa de contingut

La validesa del contingut es defineix com el grau en què els ítems de l'eina reflecteixen i són rellevants i representatius de l'univers de contingut específic de l'eina. És a dir, el que es vol determinar és si l'eina mesura tots aquells aspectes més rellevants del concepte que es vol mesurar, eliminant, a més a més, aquells que són no desitjables. Se sol establir gràcies a la definició del concepte que es vol mesurar i la identificació de les seves dimensions, a partir de revisions sistemàtiques i consultes a experts.

Validesa de criteri

Aquesta forma de validesa pretén relacionar els resultats d'una determinada eina amb una altra mesura del mateix concepte considerada i acceptada com a ideal, representativa o de referència (criteri). L'elecció del criteri és l'aspecte crític en aquest procediment validació, i l'eina es pot comparar amb un únic criteri, o amb diversos criteris. Dins d'aquest tipus de validesa, hi ha dues formes principals: la validesa predictiva i la validesa concurrent.

- La *validesa predictiva* es refereix a la capacitat d'una eina d'avaluació de predir resultats futurs en alguna activitat o en una altra avaluació del mateix constructe (amb la mateixa eina o amb una altra). La millor manera d'establir directament la validesa predictiva és realitzar un estudi de validesa a llarg termini, però es requereixen grandàries mostrals força grans per tal d'adquirir dades agregades significatives.
- La *validesa concurrent* està relacionada amb la utilització d'un criteri que està disponible en el moment de l'avaluació. És a dir, els resultats de l'eina es corresponen amb els d'un criteri ja existent establert.

Validesa de concepte o de constructe

Quan no existeix cap altra eina que mesuri el mateix que la nostra, la comparació es pot fer amb un constructe. El terme constructe fa referència al concepte teòric que es vol mesurar. Per tant, la validesa de constructe mesura fins a quin punt l'eina confirma una hipòtesi feta a priori. La validesa de constructe inclou dos components: la convergent i la discriminant.

- La *validesa convergent* fa referència al grau en què es relacionen dues eines amb constructes que s'espera que estiguin relacionats.
- La *validesa discriminant* determina fins a quin punt un fenomen discrimina altres fenòmens dissenyats per avaluar conceptes completament diferents.

A part de la validesa, existeixen també altres propietats psicomètriques relacionades al fet que l'eina estigui lliure d'error aleatori, que es presenten a continuació.

Consistència interna

La consistència interna intenta mesurar el grau en què els elements de l'eina mesuren el mateix constructe i, per tant, es considera, en general, com una evidència de l'estructura interna de l'eina. Dóna per tant, una mesura de quina és la relació que hi ha entre tots els ítems o preguntes de l'eina. La mesura de consistència interna més utilitzada és el coeficient Alpha de Cronbach, on el valor implica una correlació nul·la i el valor d'u, una correlació perfecta. Tot i que no hi ha regles absolutes sobre quins són els valors òptims de la consistència interna, la majoria d'estudis coincideix que perquè el coeficient de consistència interna es consideri acceptable, ha de ser superior a 0,69.[38,39]

Fiabilitat

Es refereix a la propietat de l'eina que fa referència al grau de repetibilitat o reproductibilitat, és a dir, en quin grau obtindrem el mateix resultat si es repeteix la mesura en unes condicions o circumstàncies similars. Existeixen diferents mitjans per provar la fiabilitat d'un instrument, en funció de les formes d'administració de l'eina:

- La *fiabilitat inter-observador* es refereix al grau d'acord entre els resultats mesurats per l'eina utilitzant dos o més observadors, i en les mateixes condicions.
- La *fiabilitat intra-observador*, també coneguda com a *fiabilitat test-retest*, descriu l'acord entre resultats quan l'eina és utilitzada pel mateix observador en dues o diverses ocasions (i en les mateixes condicions).

L'ús d'un concepte o d'un altre dependrà del tipus d'eina i dels objectius de l'avaluació.

Tot i aquestes definicions ben clares, i força utilitzades en altres àmbits diferents de l'avaluació de la recerca (com per exemple en el *Patient Reported Otcomes Measures*, PROMS), no s'ha trobat gaire evidència científica sobre la validesa dels instruments d'avaluació ex-post. En un estudi realitzat a la Universitat de Girona, es va avaluar la validesa del mètode d'avaluació ex-post a través de la consistència interna i la validesa discriminant, trobant una bona capacitat discriminatòria i una consistència per mesurar la contribució al coneixement científic i a l'impacte social.[40] Una de les limitacions que es marcaven en aquest estudi era una potencial subjectivitat en l'atribució, per part dels investigadors, de l'impacte de la seva recerca, aspecte

que podria comportar una sobreestimació dels resultats. Justament, aquest aspecte sobre fins a quin punt es pot suposar que els investigadors proporcionen respostes veraces en els qüestionaris que avaluen l'impacte de la seva recerca, està àmpliament discutit en dos estudis britànics.[14,41] En aquests estudis es conclou que, tot i la limitada evidència existent, es pot suggerir que els investigadors no exageren els impactes de la seva recerca, especialment si els resultats de l'avaluació no comporten un futur finançament o promoció. [14,41]

En aquesta tesi, ens centrarem en els conceptes de validesa de contingut, validesa discriminant i consistència interna.

OBJECTIUS I PREGUNTES DE RECERCA

Aquesta tesi doctoral ha desenvolupat diferents estratègies per mesurar la validesa dels diferents mètodes d'avaluació de la recerca en dues fases diferents del procés d'avaluació de la recerca. Per dur a terme aquest procés s'han establert dos objectius principals, un per a una avaluació ex-ante i un altre per a una avaluació ex-post:

1. Investigar, en el cas ex-ante, si anonimitzar la identitat dels investigadors i les seves institucions en la primera etapa d'avaluació dins d'un procés de revisió per consemblants de propostes de recerca modifica la valoració de l'avaluador quan aquest coneix, en una segona etapa, el nom de l'investigador/equip investigador principal, la seva experiència i la institució que representen. Concretament, es quantifica el canvi en la valoració, la seva orientació (validesa en el resultat obtingut), els factors associats i els motius que la intervingueren (validesa en la fiabilitat de la presa de decisions).
2. Comprovar, en el cas ex-post, la validesa (de contingut, discriminant i la consistència interna) d'un instrument dissenyat per mesurar els impactes observats dels projectes finançats de forma competitiva, des de la perspectiva dels investigadors. A més, mesurar els diferents nivells d'impacte que han assolit els projectes i els factors associats a aquests impactes.

Per tal de desenvolupar els objectius, s'han formulat dos grups principals de preguntes de recerca que aborden cadascun dels objectius principals. Els estudis s'han dissenyat en conseqüència. El primer grup de preguntes de recerca tracta de l'eina d'avaluació ex-ante.

RQ1. Quin és l'efecte d'anonimitzar la identitat dels investigadors i les seves institucions en la primera etapa d'avaluació d'un procés de revisió per consemblants de propostes de recerca, sobre el canvi en l'avaluació de l'avaluador un cop aquest coneix, en una segona etapa, el nom de l'investigador principal/grup de recerca, la seva experiència i la institució a la qual pertany?

RQ2. Quins factors externs afecten el procés de presa de decisions, independentment de la qualitat de la proposta que s'avalua?

Un segon grup de preguntes de recerca es relacionen amb una eina d'avaluació ex-post.

RQ3. El qüestionari d'avaluació ex post de projectes per mesurar l'impacte social de la recerca és vàlid?

RQ4. Quines són les característiques dels projectes més influents en els impactes assolits?

Capítol 2: Mètodes

Poder disposar de processos sòlids i robusts i d'instruments precisos i acurats és rellevant per a permetre una presa de decisions basada en l'evidència científica vàlida i fiable. Aquest capítol descriu els mètodes utilitzats per avaluar la validesa d'eines d'avaluació. El capítol s'estructura en dos apartats, un relacionat amb la validació d'un procés d'avaluació ex-ante, i un altre relacionat amb la validació d'una eina d'avaluació ex post i inclou tant dades qualitatives com quantitatives.

VALIDESA D'UN PROCÉS D'AVALUACIÓ EX-ANTE DE LA RECERCA

Descripció del model d'avaluació ex-ante

A continuació es descriu el model d'avaluació analitzat portat a terme a l'AQuAS per l'avaluació de convocatòries de projectes de recerca. El procés d'avaluació està basat en una avaluació per consemblants de la qualitat dels projectes on hi participen com a avaluadors diferents experts internacionals. Per aquest motiu tot el procés d'avaluació es realitza en anglès.

En un primer moment es realitza una cerca d'avaluadors a través de la literatura científica, societats científiques, editors de revistes o repositoris d'avaluadors. La selecció es porta a terme segons criteris d'experiència científica i/o metodològica, i prestigi i reconeixement en l'entorn científic i sanitari. Es considera criteri d'exclusió el fet de formar part d'equips de recerca catalans o de participar (l'avaluador o la seva institució) en algun projecte presentat a la Convocatòria. Tots els avaluadors participants han d'acceptar les normes establertes per a l'avaluació dels projectes presentats (conducta ètica, coneixement sobre el tema a avaluar, capacitat per redactar crítiques constructives i adaptació al calendari d'avaluació establert) i han de signar una declaració sobre el manteniment de la confidencialitat de la informació i l'absència de conflicte d'interessos per a cadascun dels projectes avaluats.

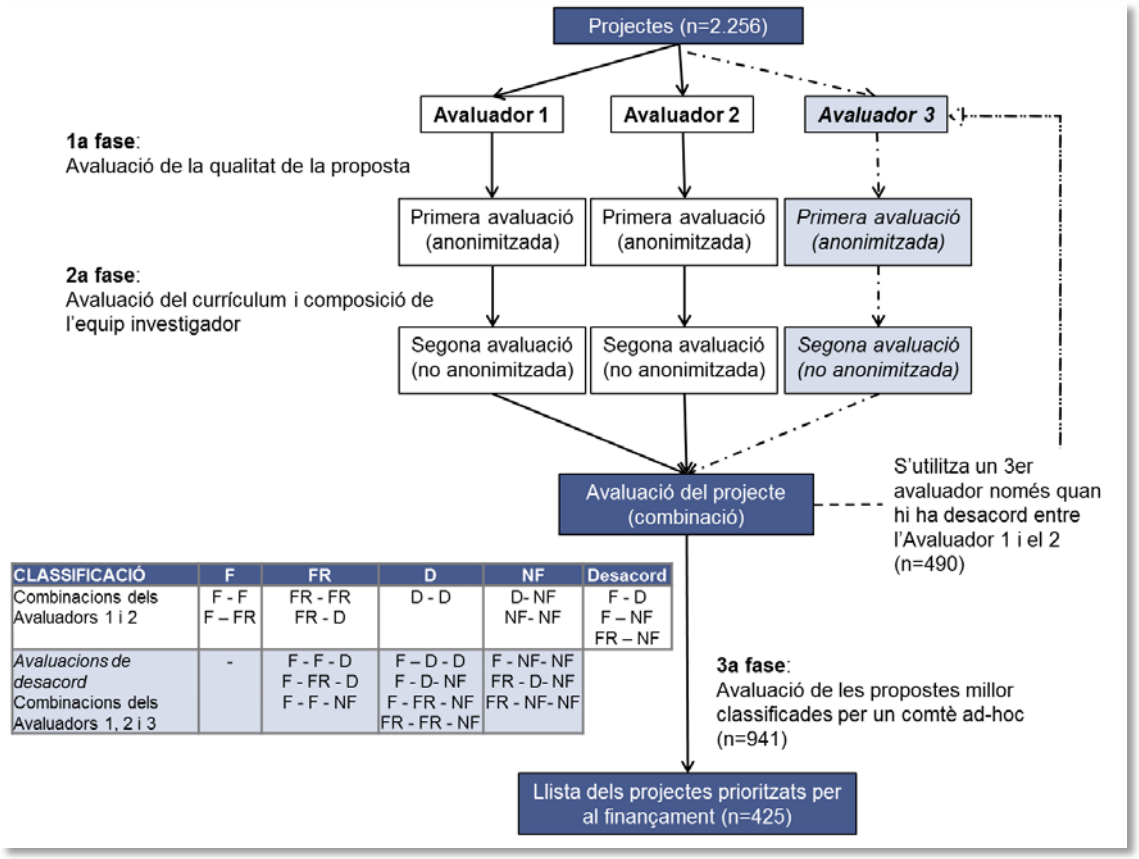
Un cop rebuda tota la informació dels projectes, aquests es classifiquen per temàtica i àrea de recerca (bàsica, clínica, epidemiològica o combinacions de les tres) segons la informació facilitada pels investigadors. Posteriorment, diferents tècnics els anonimitzen, és a dir, comproven que el projecte no inclou dades relacionades amb els investigadors vinculats al projecte, així com els detalls de les institucions en que treballen, i en cas necessari, substitueixen

aquestes dades per XXX (projectes anonimitzats). Els noms dels investigadors, els centres on treballen, els títols, volum o pàgines de les autoreferències (però no la identificació de la revista o l'any de publicació) se suprimeixen tant en el text com en la secció bibliogràfica.

El motiu de l'anonimització en una primera fase del procés d'avaluació té l'objectiu de que les avaluacions se centrin principalment en la qualitat de la proposta, tal com s'identifica en els objectius del procés de finançament. Tot i això, la realització de propostes de recerca també depèn de la capacitat dels equips de recerca per implementar adequadament les propostes i per tant, implica que la segona avaluació, de la capacitat de l'equip, és necessària, tot i que té un pes inferior.

A continuació, cada projecte s'assigna a dos avaluadors en funció de l'expertesa de l'avaluador i el tema i l'àrea de recerca del projecte. Els avaluadors actuen de forma independent (*peer review*, revisió per consemblants). El procés d'avaluació consta de tres fases (Gràfic 2).

Gràfic 2. Fases del procés d'avaluació ex-ante



F = finançable; FR = finançable amb reserves; D = dubtós; NF = no finançable

En una primera fase (anonimitzada), s'avalua el projecte anonimitzat. És a dir, l'avaluació d'aquesta primera part vol centrar-se en la rellevància, la qualitat i el rigor metodològic del projecte de recerca en funció dels aspectes següents:

- Coneixement del tema
- Adequació i especificitat d'hipòtesis, objectius i metodologia
- Rellevància científica, sanitària i social
- Factibilitat, pla de treball i cronograma presentats
- Disponibilitat de recursos i adequació del pressupost sol·licitat

Per valorar aquests aspectes els avaluadors disposen d'un qüestionari estructurat amb categories de resposta amb una escala Likert (totalment d'acord, d'acord, en desacord, totalment en desacord). Al final del qüestionari, els avaluadors han de respondre si, qualitativament, el projecte els sembla finançable (F), finançable amb reserves (FR), dubtós (D) o no finançable (NF) per al finançament. Les definicions de les diferents categories es descriuen en l'Annex 3.

Una vegada revisat el projecte anonimitzat, en la segona fase (no anonimitzada), els avaluadors disposen de la informació curricular i de nou, amb un nou qüestionari estructurat avaluen l'experiència i la trajectòria de l'investigador principal (IP), l'equip de recerca i la idoneïtat de la institució on es planteja la recerca en funció dels següents aspectes:

- Trajectòria i expertesa de l'equip investigador
- Experiència prèvia en la línia de recerca
- Composició de l'equip
- Diversitat de disciplines i centres participants

En aquest punt, els avaluadors han de donar la seva segona i definitiva avaluació qualitativa del projecte en els mateixos termes (F, FR, D o NF), tenint en compte, per tant, tant la qualitat del projecte com l'equip investigador. Això permet a l'avaluador modificar l'avaluació inicial de manera justificada. Les definicions de les diferents categories es descriuen en l'Annex 3. El temps transcorregut entre la primera i la segona avaluació és d'entre una i dues setmanes.

Si en aquest punt hi ha un desacord de dos o més nivells de categoria de les valoracions dels dos avaluadors sobre un mateix projecte, el projecte s'envia a un tercer avaluador, que el revisa i el

valora de manera independent, és a dir, sense conèixer els resultats de les avaluacions prèvies i segueix el mateix procés en dues fases.

Les dues o tres categories obtingudes a partir de les avaluacions es combinen per a cada projecte, permetent així obtenir una valoració qualitativa final que serà el criteri de referència, i que permetrà classificar els projectes en categories excloents (F, FR, D i NF). Tot aquest procés pot comprendre entre dos i quatre mesos, en funció del nombre de projectes presentats.

En aquelles convocatòries amb nombres elevats de projectes presentats, la tercera i última etapa del procés la realitza un comitè ad hoc integrat per alguns dels avaluadors participants, els quals es reuneixen durant un període de dos dies. Aquests avaluadors s'encarreguen d'avaluar les millors propostes qualificades en la segona etapa (generalment les del criteri de referència F, i de sovint també les FR). El comitè redacta una llista dels projectes prioritzats per al seu finançament, tenint en compte l'import dels fons disponibles, que es presenta a la comissió consultiva de l'entitat finançadora. En aquelles convocatòries amb nombres reduïts de projectes (menys de 10), aquesta tercera part, si és necessària, es discuteix en línea.

En aquesta tesi només es tindrà en compte els resultats fins a la segona fase, primer perquè aquesta és la part homogènia en el procés d'avaluació utilitzat en diferents convocatòries (si considerem que en les avaluacions de convocatòries amb un nombre petit de propostes presentades no es considera necessari realitzar la tercera fase amb el comité ad-hoc); i segon, perquè les dues primeres fases les considerem un punt clau, ja que és on es realitza la primera selecció i per tant, es decideix aquelles propostes que en cas d'existir prou finançament podrien rebre'l.

Disseny i població de l'estudi

S'ha realitzat un estudi retrospectiu observacional del procés de revisió per consemblants realitzat per avaluar diferents projectes de recerca en ciències de la salut. La mostra ha estat formada per tots els projectes presentats en 14 convocatòries portades a terme entre el 2001 i el 2014 (n=2.256) gestionades per l'AQuAS. Aquests projectes van ser avaluats per 1.475 avaluadors (amb una mitjana de 2,2 d'avaluacions per avaluador). En total, es van realitzar 5.002 avaluacions.

Influència del currículum de l'equip investigador en el canvi de l'avaluació

Per tal d'avaluar la validesa en el resultat obtingut, és a dir, el pes de la qualitat científica del projecte, objectiu principal dels finançadors, s'ha analitzat la influència del currículum de l'equip investigador en el canvi d'avaluació.

La variable principal d'estudi és el canvi realitzat entre la primera avaluació (projecte anonimitzat) i la segona avaluació (projecte no anonimitzat) per a cada avaluació i per a cada avaluador (intravaluador). La variable s'ha classificat en una escala ordinal de 0 = no canvi, 1 = millora (canvi a una categoria superior) i 2 = empitjora (canvi a una categoria inferior).

Per tal d'analitzar la influència del currículum de l'equip investigador en el canvi de valoració, es va analitzar la relació entre la variable principal i l'adequació de l'IP/equip investigador/institució de l'IP amb una prova de Chi-quadrat. Aquesta variable s'ha extret de la segona fase del model d'avaluació, on els avaluadors valoren amb una pregunta la idoneïtat de l'IP i l'equip de recerca per a dur a terme el projecte en una de les quatre categories Likert següents: molt d'acord, d'acord, en desacord o totalment en desacord. Aquesta variable va ser dicotomitzada per a aquesta tesi. L'acord entre les primeres avaluacions (anonimitzades) i les segones (no anonimitzades) es va calcular utilitzant l'estadística kappa ponderada (k). El coeficient kappa pren valors entre -1 i +1; com més proper és a 1, més gran és el grau de concordança, mentre que el valor zero reflecteix que la concordança observada és precisament la que s'espera a causa exclusivament de l'atzar, valors superiors a 0,60 seran considerats amb un grau de concordança bo.[42]

Motius que han influït en el canvi de l'avaluació segons l'opinió dels investigadors

S'ha analitzat la fiabilitat en termes de la qualitat de la presa de decisions, determinant els motius que influencien el canvi de valoració.

Per analitzar aquests motius que han influït indirectament en un canvi en la segona avaluació (no anonimitzada) es va dur a terme una anàlisi qualitativa de contingut. Atès que el qüestionari d'avaluació no té un camp específic per documentar els motius d'un canvi, s'ha analitzat el camp obert per a comentaris addicionals inclosos en el formulari d'avaluació emplenat en la segona fase. Els únics casos que s'han examinat són aquells en què l'avaluador realitza un canvi entre l'avaluació del projecte anonimitzat (primera fase) i l'avaluació del projecte no anonimitzat (segona fase) i que a més, s'havia emplenat el camp de comentaris addicionals.

L'anàlisi del contingut s'ha dut a terme d'una manera inductiva en dues dimensions: motius del canvi i naturalesa del canvi. En primer lloc, de tots els comentaris escrits pels avaluadors s'ha obtingut una comprensió global dels mateixos i s'han extret les categories i subcategories inicials i d'aquesta manera s'han classificat tots els comentaris en aquestes categories i subcategories. En segon lloc, s'ha determinat si les justificacions són de naturalesa positiva, negativa o neutre. Una justificació s'ha definit com positiva quan l'avaluador descriu aspectes rellevants i fortaleses del projecte o de l'equip investigador; una justificació negativa significa que les característiques avaluades no semblen adequades o estan absents i per tant s'indica debilitats; i per últim, s'ha considerat les justificacions com a neutres quan no es destaquen ni fortaleses ni debilitats. Els resultats han estat triangulats per dos codificadors, que han arribat a un consens i han discutit aquells casos classificats com a 'dubtosos'.

Factors associats amb un canvi substancial de l'avaluació

S'ha analitzat la fiabilitat del procés determinant els factors o biaixos que es poden associar a la presa de decisions.

S'ha utilitzat un model de regressió logística multinomial ajustat per identificar els factors associats al canvi. La categoria 'no canvi' es va utilitzar com a categoria de referència. La Reducció Relativa del Risc (RRR) s'ha calculat, doncs, per les categories de 'millora' i 'empitjora'. En l'anàlisi de dades s'han inclòs diverses covariables com a predictors, en dos nivells: a nivell de projecte i a nivell d'avaluador. En la següent taula es descriuen les diferents variables.

Taula 1. Definició i categorització de les variables d'estudi en l'avaluació ex-ante

Nivell	Variable	Definició	Categorització
Projecte	Any d'edició	Anualitat en la que s'ha obert la convocatòria	2001-2014
	Àrea de recerca del projecte	Tipus de recerca del projecte segons l'equip investigador	Bàsica Clínica Epidemiològica Combinacions de les anteriors
	Quantitat de subvenció sol·licitada	Suma màxima de la subvenció sol·licitada	<100.000€ 100.000–199.999€ 200.000–299.999€ ≥300,000
	Sexe de l'IP	El sexe del sol·licitant s'ha obtingut a partir dels formularis completats pels mateixos	Dona Home
	Edat de l'IP	L'edat del sol·licitant en el moment de presentar el projecte i calculada a partir de la data de naixement que presenten en el formulari	≤40 >40
	Valoració de l'experiència adequada de l'equip investigador	Extret de la resposta que els avaluadors fan a la pregunta: *The skills and experience of the research team is adequate to carry out the proposal, en la segona fase de l'avaluació*	Adequada: Completament d'acord/ d'acord No adequada: En desacord/ completament en desacord
Avaluador	Sexe de l'avaluador	Obtingut a partir dels formularis d'acceptació de l'avaluador. Si el sexe no s'ha identificat específicament, s'ha realitzat una cerca manual (utilitzant llocs web o adreces corresponents). En particular, els noms asiàtics han estat difícils d'assignar a un sexe específic, de manera que en aquets casos la variable ha quedat en blanc (i no entra en l'anàlisi)	Dona Home
	Regió mundial de l'avaluador	Regió mundial de la institució on treballa l'avaluador	Europa Nordamèrica Altres
	H-index de l'avaluador	L'índex H s'ha determinat per a cada avaluador utilitzant Web of Science i tenint en compte l'any de l'avaluació. El valor de H és igual al nombre d'articles de l'avaluador (N) que tenen N o més cites. S'utilitza aquest valor com a proxy de l'experiència de l'avaluador	≤15 >15

IP: investigador principal

El càlcul del grau d'associació s'ha realitzat amb l'estadístic Reducció Relativa del Risc (RRR) que mira la diferència de risc entre el grup que 'millora' en el canvi amb el que 'empitjora' en el canvi

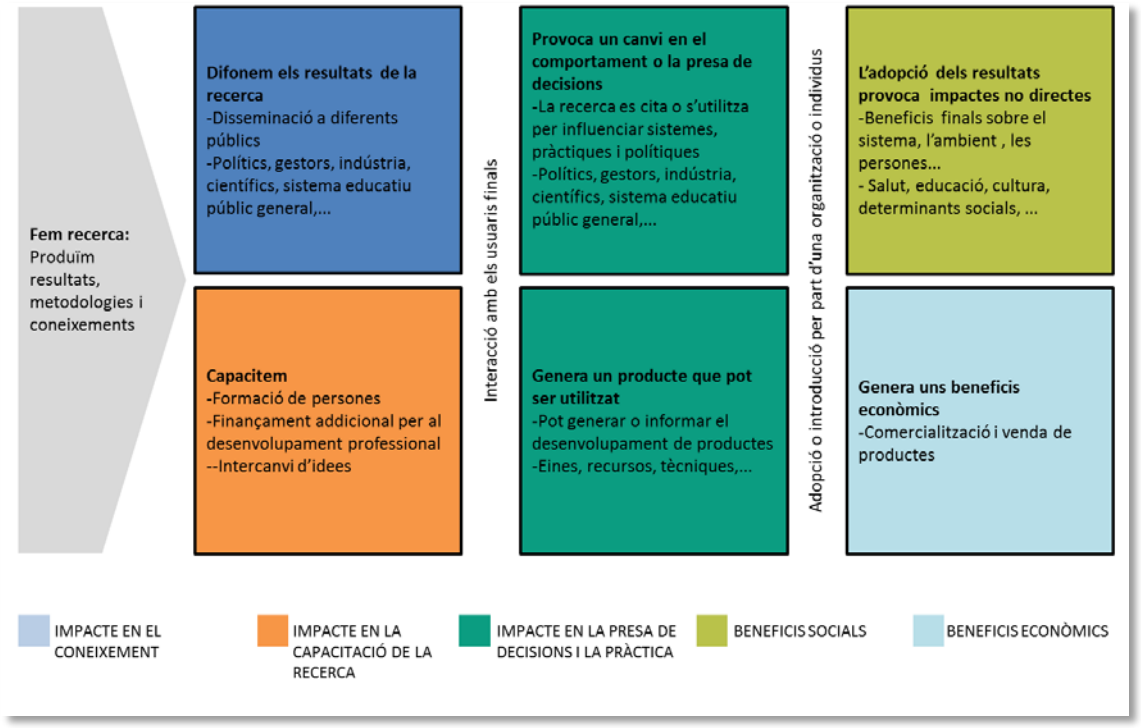
respecte al grup que no canvia la valoració. La significació estadística de totes les anàlisis s'ha fixat en p≤0.05. Les estadístiques es van calcular utilitzant SPSS18.

VALIDESA D'UNA EINA D'AVALUACIÓ EX-POST DE LA RECERCA

Descripció del model d'avaluació ex-post

L'avaluació ex-post s'ha centrat en la mesura de l'impacte de la recerca portada a terme per projectes de recerca finançats competitivament. En aquesta part de l'estudi s'ha definit l'impacte de la recerca com els beneficis reals experimentats que es deriven dels resultats de la recerca aplicables a una àmplia gamma de beneficiaris, incloent-hi a les persones, les organitzacions, les comunitats, les regions o altres entitats. Per a l'avaluació de l'impacte de la recerca s'ha utilitzat el model *Payback*[14], àmpliament emprat per analitzar l'impacte de la recerca en diverses disciplines,[43] tot i que on més ha estat provat és en la recerca en ciències de la salut i de la vida, així com àmpliament utilitzat per a l'avaluació de l'impacte de projectes de recerca. El *Payback* consta de tres elements. El primer, un model lògic que identifica una multiplicitat d'elements que formen part del procés de recerca que van des de la conceptualització de la recerca fins al seu impacte. El segon, dues 'interfícies' que volen mostrar la interacció entre els investigadors i els usuaris potencials de la recerca (configuració de l'agenda de recerca i difusió de la recerca). El tercer element són cinc nivells d'impacte: *impacte en el coneixement* (representada per publicacions científiques o difusió a públics no científics); *impacte en la capacitació de la recerca* (formació de recerca, noves col·laboracions, assegurament de fons addicionals o millora de les infraestructures); *impacte en la presa de decisions i la pràctica* (recerca utilitzada com evidència en la presa de decisions, en un ampli ventall de circumstàncies, i en el desenvolupament de pràctiques); *beneficis socials* (aplicació de la recerca dins del sector de la disciplina); i *beneficis econòmics* (explotació comercial o ocupació). En el gràfic 3 s'ha intentat explicar el que poden ser els diferents efectes de l'impacte de la recerca, per tal d'ajudar a entendre la multidimensionalitat del concepte d'impacte.

Gràfic 3. Efectes de l'impacte de la recerca i nivells d'impacte



Disseny i població de l'estudi

S'ha realitzat un estudi transversal per avaluar l'impacte de la recerca realitzada a la Universitat Oberta de Catalunya (UOC). La recerca de la UOC inclou quatre àrees de coneixement: arts i humanitats, ciències socials, ciències de la salut, i tecnologies de la informació i la comunicació (TIC). Els temes de recerca tractats inclouen una gran diversitat com ara 'identitat, cultura, art i societat'; 'tecnologia i acció social'; 'globalització, pluralisme legal i drets humans'; 'impostos, relacions laborals i beneficis socials'; 'Internet, tecnologies digitals i mitjans de comunicació'; 'gestió, sistemes i serveis en informació i comunicacions' i 'eHealth'. El Comitè d'Ètica de la UOC va aprovar l'estudi.

S'ha volgut aprofitar la visió de la UOC en quant a avançar sobre el paper social i l'impacte de la recerca realitzada pels seus investigadors, amb diferents iniciatives pioneres que ha adoptat al seu actual pla estratègic, i que inclouen la promoció del coneixement obert, una mesura específica relacionada amb l'impacte social de la recerca, i el desenvolupament d'una pla d'acció de tota la institució per incorporar l'impacte en l'avaluació de la recerca.[44] La UOC també està implicada en implementar en processos d'avaluació institucional les recomanacions de la

Declaració de San Francisco DORA.[45] A més, la UOC també està al capdavant de la consecució dels objectius de desenvolupament sostenible (ODS) de l'Agenda 2030 de les Nacions Unides, després de ser seleccionada per l'Associació Internacional d'Universitats com una de les 16 universitats de tot el món que lideren aquests ODS.[46]

La població estudiada incloïa tots els IPs que disposaven d'almenys un projecte finançat competitivament (públic o privat), el qual havia d'haver finalitzat no més tard del 2017 (n = 159). La UOC , al 2017 tenia un total de 436 investigadors.

Impactes percebuts i factors associats

Es va dissenyar un qüestionari en línia per a ser completat pels IPs dels projectes amb l'objectiu de determinar retrospectivament els impactes percebuts que s'atribuïen directament als projectes. Els IPs havien d'escollir una recerca (un projecte únic o un grup de projectes relacionats en una temàtica conjunta) sorgida de la seva àmplia cartera de treballs, que ja estigués finalitzat i que tingués algun impacte evident i real en qualsevol dels nivells en el moment de contestar el qüestionari. Per al desenvolupament d'indicadors es va tenir en compte l'experiència prèvia de l'equip i treballs publicats a la literatura científica.[14,47] El qüestionari es va estructurar entorn de la categorització multidimensional dels impactes del model Payback.

El qüestionari va incloure quatre seccions. La primera secció va registrar informació sobre els IPs, incloent-hi el sexe, l'edat i el nombre d'anys d'experiència en recerca. El segon, se centrava en la naturalesa del mateix projecte, incloent-hi la disciplina, el tema principal de la recerca, els conductors originals de la recerca, la interacció amb els usuaris potencials de la recerca i els organismes de finançament. En la tercera secció es van abordar les percepcions de l'impacte de la recerca en cadascuna de les cinc categories d'impacte esmentades anteriorment. A la taula 2 s'inclouen les cinc dimensions d'impacte amb les diferents subcategories d'impacte (ítems). Els ítems del qüestionari incloïen preguntes amb respostes dicotòmiques (sí/no) i preguntes obertes addicionals per a afegir una breu descripció dels impactes percebuts. Una última secció incloïa qüestions generals, una de les quals pretenia captar altres impactes rellevants que podrien no estar dins dels ítems del qüestionari i una pregunta final on es volia determinar una avaluació (en percentatge) de la contribució/atribució de la recerca a cadascuna de les cinc categories d'impacte. Els enquestats havien de decidir el nivell de contribució/atribució dels impactes segons tres categories de resposta: contribució limitada de la recerca, entre 1 i un 30%; contribució moderada, del 40-60%; i contribució significativa del 70-100%.

Taula 2. Dimensions i subcategories d'impacte

Dimensió	Indicador
Impacte en el coneixement i la seva difusió	Resums de congressos
	Articles científics
	Llibres o capitols de llibres
	Materials educatius
	Presentacions al públic general
	Presentació als participants de la recerca
	Mencions en els mitjans de comunicació
	Xarxes socials
	Blogs influents
	Presentacions en concerts, enregistraments o sales de música
Impacte en la capacitació	Formació d'estudiants de doctorat
	Formació d'estudiants de màster
	Formació d'estudiants de grau
	Noves col·laboracions a nivell estatal
	Noves col·laboracions a nivell internacional
	Noves xarxes acadèmiques
	Finançament addicional per a nous projectes de recerca
	Finançament addicional pel grup de recerca
	Recerca o mètodes utilitzats per altres investigadors
	Millores en les infraestructures
Impacte en la presa de decisions i la pràctica	Informant en processos de discussions, assessories o debats per a la presa de decisions
	Informant en la formulació de normes, iniciatives polítiques o recomanacions per a reguladors
	Contribució en el disseny, planificació i gestió de serveis i prioritats
	En la implementació, adopció o producció de pràctiques fora i dintre del món professional
	Influint el comportament dels professionals o d'altres
	Influint els sistemes d'educació i les avaluacions curriculars
Beneficis socials	En la salut
	En la qualitat de vida
	En determinants socials I culturals
	En determinants ambientals
	Acceptació
	Accessibilitat
	Continuïtat
	Efectivitat i eficiència
	Seguretat
	Benestar I beneficis socials
	Preservació del patrimoni
	Competitivitat i desenvolupament d'estímuls
Beneficis econòmics	Obtenint alguna patent /protecció intel·lectual
	Rebent ingressos per *royalties*, contractes amb la indústria i/o prestació de serveis
	En la creació d'empreses spin-offs i/o start-ups
	En la creació d'acord de Material Transfer Agreements (MTS)
	Portant al mercat innovacions, productes o dispositius produïts pel sector privat
	Creant nous llocs de treball
	Beneficis econòmics cap a la societat

Els IPs van ser contactats per correu electrònic i van ser informats dels objectius de l'estudi, assegurant que les dades serien tractades confidencialment. Els IPs varen rebre dos recordatoris, també per correu electrònic. Els IPs implicats en més d'un projecte se'ls va demanar de seleccionar-ne un projecte únic o un grup de projectes relacionats en una temàtica conjunta, per reduir el temps necessari per completar l'enquesta i, per tant, augmentar la taxa de resposta.

Es va realitzar un exercici de control de qualitat abans de l'anàlisi de dades, comprovant la classificació correcta dels diferents impactes comparant les respostes sí/no amb la informació proporcionada en les preguntes addicionals obertes. No es van necessitar alteracions després d'aquestes comparacions. Els resultats dels qüestionaris varen proporcionar una mesura del nombre de projectes de recerca que contribueixen a un tipus particular d'impacte; per tant, per estimar cada nivell d'impacte, es va calcular la freqüència de la seva aparició en relació amb el nombre de projectes. Per analitzar els factors associats, es va utilitzar una prova de Chi quadrat per analitzar les diferències entre les diferents categories d'impacte i les característiques descriptives dels IPs i dels projectes.

Validesa de l'instrument

Abans del treball de camp, i com a part de la validesa de contingut, es van revisar els resums de 72 casos d'impacte (dos per àrea de coneixement) extrets de la base de dades del REF2014. El REF2014[48] és un exercici a escala nacional que avalua l'impacte de la recerca universitària al Regne Unit més enllà de l'àmbit acadèmic i compta amb una base de dades pública amb més de sis mil casos d'impacte de la recerca, agrupats en 34 àrees de coneixement. Els estudis de cas es van seleccionar aleatòriament i es van classificar els impactes trobats en cadascun dels casos en els diferents ítems del qüestionari per tal de comparar si es podia realitzar una correcta classificació dels impactes en les diferents àrees de coneixement i valorar la inclusió de nous ítems. Aquesta revisió va ajudar a reformular i afegir preguntes, especialment en les seccions d'impacte en la presa informada de decisions i en la de beneficis socials.

Per a la validesa de l'instrument, primer de tot, es va avaluar la consistència interna. Es va avaluar tant a escala de tot el qüestionari com per a cada una de les dimensions utilitzant l'alfa de Cronbach (α). L'alfa de Cronbach pren valors entre 0 i 1 i es considera acceptable quan aconsegueix valors entre 0,70 i 0,95.[49]

Per a valorar la validesa de contingut, el qüestionari va ser enviat a experts en diverses disciplines. Els experts havien de valorar la rellevància de cada ítem del qüestionari en funció de la definició de l'impacte utilitzada en l'estudi. Els experts van classificar cada pregunta en una escala de 4 categories (0 = 'no rellevant', 1 = 'lleugerament rellevant', 2 = 'bastant rellevant', 3 = 'molt rellevant'). També se'ls va demanar si creien que els diferents ítems cobrien tots els aspectes importants o si creien que faltava algun ítem per afegir. Amb les puntuacions del qüestionari es va crear un índex d'opinió d'experts que calculava la validesa de l'ítem. Aquest índex es va calcular dividint el nombre d'experts que proporcionen una puntuació de 2 o de 3, entre el nombre total de respostes. A causa de la diversitat de disciplines i temes avaluats, es van calcular diferents índexs considerant d'una banda a tots els experts, així com diferenciant els experts de cada disciplina. Aquest índex es va considerar acceptable si el nivell d'aprovació era ≥ 0,5.

Per a avaluar la validesa discriminant, es va estudiar la distribució de les respostes en termes d'efectes sòl i sostre. L'efecte sòl inclou aquells ítems on cap dels IPs percep un impacte, mentre que l'efecte sostre inclou aquells ítems on tots els IPS descriuen l'impacte. Els efectes sòl i sostre redueixen la capacitat discriminatòria del qüestionari.

Totes les dades es van introduir en el programa estadístic SPSS18, i es va considerar el nivell de significació ≤0,05 per a totes les proves.

Capítol 3: Resultats

Aquest capítol descriu els resultats de l'estudi per a cadascun dels dos exemples descrits en el capítol 2. En primer lloc, els resultats s'examinen des de la perspectiva del procés d'avaluació ex-ante, que té per objectiu la selecció de projectes en una convocatòria de recerca. En segon lloc, els resultats s'examinen a partir del desenvolupament d'una eina d'avaluació ex-post, dedicada a recollir els impactes de la recerca percebuts pels investigadors. En l'Annex 1 i 2 s'inclouen les dues publicacions que deriven dels resultats de cadascun dels apartats. Ambdós publicats a la revista Research Evaluation, de l'editorial Oxford University Press, i situada al Q1 de Scimago, amb un factor d'impacte als 3 anys de 3,4.

VALIDESA D'UN PROCÉS D'AVALUACIÓ EX-ANTE DE LA RECERCA

Descriptiu

Al llarg de les diferents edicions analitzades, s'han examinat 2.256 projectes presentats i acceptats per a ser avaluats, liderats per 1.640 IPs. En l'avaluació d'aquests han participat 1.475 avaluadors diferents, experts internacionals, que han realitzat un total de 5.002 avaluacions, tenint en compte que les avaluacions es realitzen per consemblants i, en algun cas, amb tres avaluadors independents quan existeixen discordances entre les avaluacions. Es va necessitar un tercer avaluador per resoldre discordances en 490 casos (9,8%). La majoria dels projectes són de recerca bàsica i liderats per homes de més de 40 anys. El perfil majoritari dels avaluadors és home i pertanyent a una institució europea. En la taula 3 es mostra la descripció dels projectes i de les avaluacions.

Taula 3. Descripció i característiques dels projectes i dels avaluadors analitzats

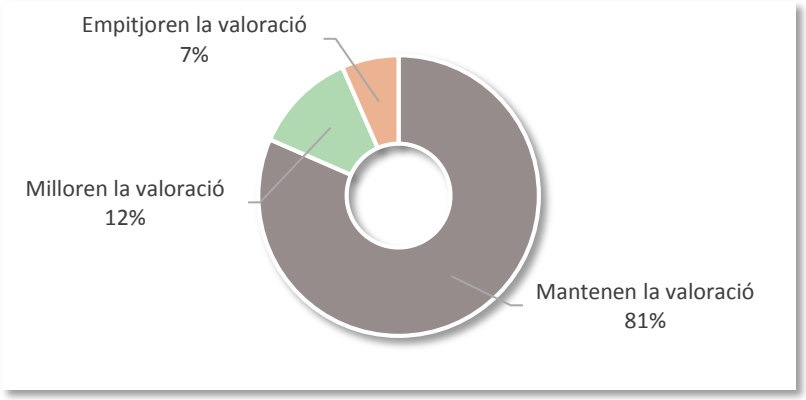
Variable	Categoria	Projectes/ IP/ Avaluadors		Avaluacions[a]	
		N	%	N	%
Any d'edició	*2001*	107	4,74	214	4,28
	2002	80	3,55	160	3,20
	2003	105	4,65	237	4,74
	2004	260	11,52	586	11,72
	2005	200	8,87	451	9,02
	2006	81	3,59	185	3,70
	2007	159	7,05	356	7,12
	2008	151	6,69	339	6,78
	2009	244	10,82	541	10,82
	2010	100	4,43	222	4,44
	2011	205	9,09	457	9,14
	2012	317	14,05	703	14,05
	2013	147	6,52	326	6,52
	2014	100	4,43	225	4,50
Àrea de recerca del projecte	*Bàsica*	1.086	48,14	2.388	47,74
	Clínica	610	27,04	1.358	27,15
	Epidemiològica	140	6,21	306	6,12
	Combinacions	420	18,62	950	18,99
Quantitat de subvenció sol·licitada	*<100.000€*	214	9,49	473	9,46
	100.000–199.999€	1.132	50,18	2.507	50,12
	200.000–299.999€	419	18,57	933	18,65
	≥300,000€	491	21,76	1.089	21,77
Sexe de l'IP	*Dona*	583	33,55	1.633	33,25
	Home	1.057	64,45	3.339	66,75
Edat de l'IP	*≤40*	340	20,73	799	15,97
	>40	1.199	73,11	3.976	79,49
	Sense informació	101	6,16	227	4,59
Experiència adequada de l'equip investigador	*Adequada*	-	-	4.316	86,29
	No adequada	-	-	613	12,26
	Sense informació	-	-	73	1,46
Sexe de l'avaluador	*Dona*	304	20,61	933	18,65
	Home	1160	78,64	1.024	20,57
	Sense informació	11	0,75	45	0,90
Regió mundial de l'avaluador	*Europa*	811	54,98	2.844	56,86
	Nordamerica	501	33,97	1.698	33,95
	Altres	163	11,05	460	9,20
Índex H de l'avaluador	*≤15*	735	49,83	2.284	45,66
	>15	699	47,39	2.584	51,66
	Sense informació	41	2,78	134	2,68
Resultat de l'avaluació – Primera avaluació	*F*	-	-	1.284	25,67
	FR	-	-	1.695	33,89
	D	-	-	1.365	27,29
	NF	-	-	651	13,01
	Sense informació	-	-	7	0,14
Resultat de l'avaluació – Segona avaluació	*F*	-	-	1.374	27,47
	FR	-	-	1.752	35,03
	D	-	-	1.256	25,11
	NF	-	-	620	12,40

[a] Cada projecte pot tenir entre 2-3 avaluacions; IP: Investigador Principal; F: finançable; FR: Finançable amb reserves; D: Dubtós; NF: No finançable

Influència del currículum de l'equip investigador en el canvi de l'avaluació

En el gràfic 4 es mostra el percentatge de canvi o no canvi de la valoració de l'avaluador després de conèixer l'IP/equip/institució. S'observa que un 81,5%, mantenen la valoració després de conèixer el CV dels investigadors. Tanmateix, la informació sobre l'equip investigador que realitzarà el projecte fa modificar la valoració prèvia de l'avaluador en un 18,5% (n=922) dels casos, tant per millorar la valoració (11,9%, n=594) com per empitjorar-la (6,6%, n=328).

Gràfic 4. Percentatge de no canvi i canvi ('millora' i 'empitjora') entre les dues valoracions (primera i segona)



En la taula 4 s'observa el percentatge de correlació de les dues valoracions (primera i segona) segons cada categoria qualitativa. En les diagonals de la taula es mostren les coincidències de valoracions, mentre que les franges verdes mostren les millores produïdes en cada una d'elles i les franges vermelles l'empitjorament en la valoració. Els percentatges més elevats de canvi estan entre categories qualitatives contínues i valoracions intermitges, i en cap cas canvia a un valor de més de dues categories.

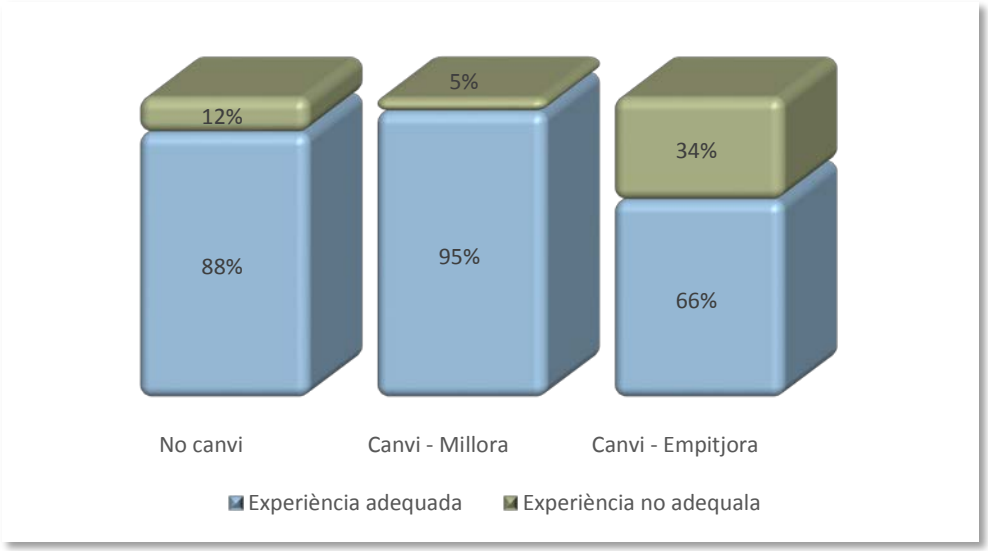
Taula 4. Percentatge de correlació de les dues valoracions (primera i segona) en les diferents categories qualitatives (n=4.995)*

		Segona avaluació				
		F	FR	D	NF	Total
Primera avaluació	F	1.158 (90,2%)	120 (9,4%)	6 (0,5%)	0 (0,0%)	**1.284**
	FR	203 (12,0%)	1.368 (80,7%)	111 (6,6%)	13 (0,8%)	**1.695**
	D	10 (0,7%)	257 (18,8%)	1.020 (74,7%)	78 (5,7%)	**1.365**
	NF	0 (0,0%)	7 (1,1%)	117 (18,0%)	527 (81,0%)	**651**
	Total	**1.371 (27,5%)**	**1.752 (35,1%)**	**1.254 (25,1%)**	**618 (12,4%)**	**4.995**

*Hi ha un 0,14% (n=7) de pèrdues, F: finançable, FR: finançable amb reserves, D: dubtós; NF: no finançable. En color verd, les valoracions que milloren en la segona avaluació (després de mirar la part curricular). En color vermell, les avaluacions que empitjoren en la segona valoració (després de mirar la part curricular).

L'associació entre l'experiència adequada de l'IP/equip de recerca i el canvi en la segona avaluació va ser estadísticament significativa (p <0,001). El canvi va ser de 'millora' quan hi va haver una valoració positiva de l'experiència de l'IP/equip de recerca, mentre que va 'empitjorar', quan l'experiència de l'equip investigador es va valorar negativa (Gràfic 5). Tant els percentatges de les categories (F, FR, D i NF) de la primera avaluació (projecte anonimitzat), com de la segona avaluació (amb informació de l'IP/equip de recerca), van ser semblants. L'estadístic de Kappa ponderat va indicar un acord molt bo (k=0,75) entre les primeres i segones avaluacions.

Gràfic 5. Percentatge de canvi segons l'experiència adeguada de l'IP/equip de recerca



Motius que han influït en el canvi de l'avaluació segons l'opinió dels investigadors

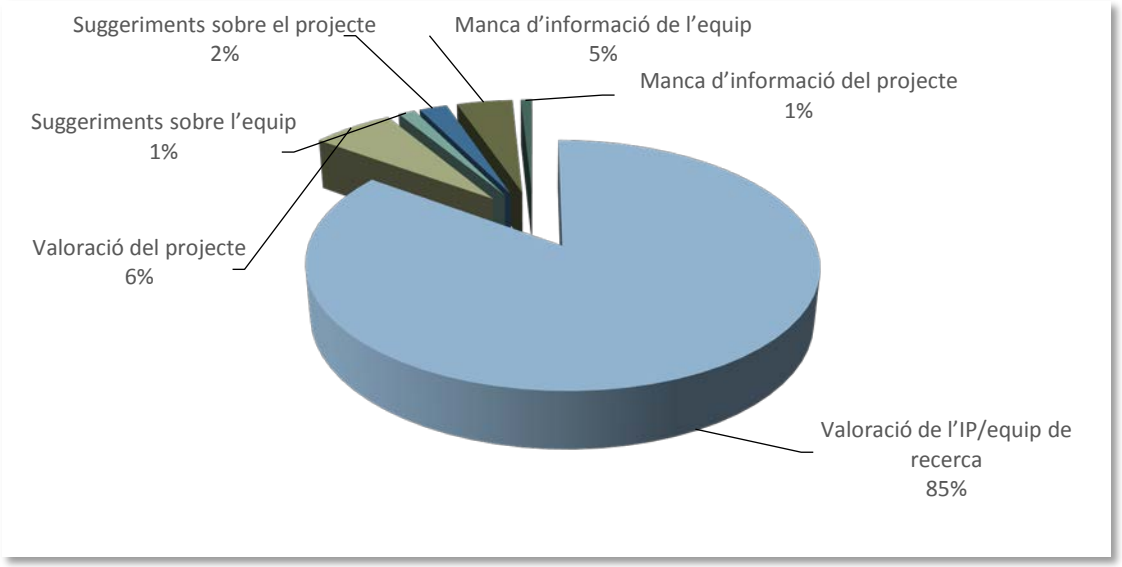
Per analitzar el perquè del canvi de valoració, s'ha realitzat una anàlisi de contingut del camp obert corresponent a l'equip investigador, on els avaluadors justifiquen les seves valoracions sobre l'equip en aquells casos en què hi ha hagut un canvi de valoració. Tot i que pot haver-hi més d'una justificació, s'ha categoritzat el motiu principal que l'avaluador ha manifestat en una única categoria. Del total de casos on hi ha hagut un canvi de valoració (n=922), la meitat dels casos ha emplenat el camp de comentaris (n=461). En un 2,5% (n=23) d'aquests casos, els comentaris són inconsistents amb el canvi de valoració, és a dir, l'avaluador millora la valoració, tot i que els comentaris són desfavorables (n=18), o bé l'empitjora, tot i que els comentaris són favorables (n=5), sense una justificació clara del canvi. Tal com s'ha explicat a l'apartat de mètodes, la justificació s'ha considerat 'positiva', quan es presenten les fortaleses de l'IP/equip de recerca o el projecte, i 'negatives', quan la valoració implicava les seves debilitats. A més, un 4% dels comentaris estan relacionats amb suggeriments o manca d'informació de l'equip o del projecte, tot i que no es comenta explícitament que sigui el motiu del canvi i, per tant, no fan cap valoració que justifiqui el canvi. Cal tenir en compte que en la pregunta no es sol·licita una justificació explícita del canvi de valoració, sinó que tan sols es demanen comentaris sobre la valoració de l'equip investigador. Les justificacions dels avaluadors van ser d'una llargada d'entre 2 i 452 paraules (de mediana la llargada és de 48 paraules, i de mitjana 71).

Els motius que sorgeixen dels comentaris dels avaluadors es varen dividir en 3 grups principals: una valoració, un suggeriment, o un comentari sobre la manca d'informació relacionada amb l'IP/equip de recerca o amb el projecte. Els comentaris que defineixen el primer grup (valoració) estan relacionats amb 1) si l'IP/equip de recerca té suficient experiència per a dur a terme el projecte, mesurat principalment amb les publicacions científiques; 2) amb la col·laboració multiinstitucional, relacionada tant amb la necessitat d'incorporar més centres participants (per exemple, per recollir suficients mostres), com amb aspectes sobre com s'han de coordinar els múltiples centres; 3) amb la composició i la naturalesa multidisciplinària de l'equip, descrita en relació amb l'experiència i especialització dels seus membres; i 4) amb l'avaluació del projecte, relacionat amb els seus punts forts i febles, com ara la rellevància, la qualitat o la novetat, característiques que eren en realitat part de l'avaluació de la primera avaluació. El segon grup (suggeriment) inclou 1) suggeriments sobre l'equip de recerca, relacionats principalment amb la necessitat de coneixements addicionals i 2) suggeriments sobre el projecte, incloent-hi una àmplia varietat d'aspectes com el pressupost, la hipòtesi, la planificació, etc. Finalment, el tercer grup, definida com a manca d'informació, es relaciona quan la justificació del revisor descriu que

la proposta no té suficient informació per avaluar un aspecte específic (de l'equip de recerca o de la proposta).

D'altra banda, aquestes categories mostren una distribució desigual, com s'observa en el gràfic 6. Tal com era d'esperar, les valoracions sobre l'IP/equip de recerca són els principals motius per fer el canvi de valoració.

Gràfic 6. Distribució dels motius per fer el canvi en l'avaluació



En la taula 5 es detallen els motius que influencien el canvi segons les opinions dels avaluadors, tant per canvis positius com negatius.

Taula 5. Descripció dels motius que influencien el canvi de valoració segons les opinions dels avaluadors

Motius del canvi	Classificació	Tipus de canvi	*Exemples*
Valoració de l'IP/equip de recerca	Experiència científica	Positiu	*The experience of the researchers changed my mind regarding the unclear feasibility of the study. The study will be quite feasible. Even though some concerns regarding study design and data analysis remain, the study has potential. [ID 17]*
		Negatiu	*This project would be done by three people with rather varying degrees of expertise. Having reviewed the C's it is doubtful to the reviewer whether this group may fulfil the ambitious goals set forth in the research plan. [ID 624]*
	Multicentrisme i col·laboració	Positiu	*The different centres involved in the work for this proposal have been collaborating in the past with good results. This can guarantee the success of proposal although the ambitions are very high. [ID 223]*
		Negatiu	*Given the multicenter structure, it is not sufficiently clear how this complicated effort will be coordinated. [ID 562]*
	Composició I multidisciplinarietat	Positiu	*The multidisciplinary team is indeed composed of researchers having different backgrounds, expertise, and specialization. Most important, however, is the difference in scientific merits. [ID 667]*
		Negatiu	*My initial enthusiasm for this proposal is somewhat diminished by the lack of an expert in cell biology on the research team. The team is very strong in protein biochemistry, but not biological studies that are important for this project. [ID 1170]*
Valoració de la proposta	Projecte	Positiu	*The applicant refers to important preliminary data that were not available in the first part of the evaluation; I have therefore reconsidered my evaluation in a positive way. [ID 2751]*
		Negatiu	*Even though in the past this group has published studies on melanoma patients, in this project no clinical studies are scheduled. This makes the project not particularly interesting. [ID 939]*
Suggeriments sobre l'equip de recerca	Composició I multidisciplinarietat	Positiu	*The design and statistics of the protocol are flawed and suggest that a statistician would be a worthwhile member of the team. [ID 1592]*
		Negatiu	*The team should have exploited their background basic immunological studies on [….] Furthermore, it is not clear who is doing what and where are the meeting points between three discipline [ID 118]*

Taula 5. Descripció dels motius que influencien el canvi de valoració segons les opinions dels avaluadors (continuació)

Motius del canvi	Classificació	Tipus de canvi	*Exemples*
Suggeriments sobre la proposta	Projecte	Positiu	*I advise the team of the project to focus on one or two tumours seen mostly in their centre e.g. neuroblastoma (as most publications are on neuroblastoma and few are on sarcomas) and to focus only on one or two translational research issues (e.g. Thyrosine hydroxylase…) and not to try to do many molecular techniques in the same project. [ID 1014]*
		Negatiu	*The proposal is planned for three years, but with well-equipped and experienced team, such experiment could be done in maximum 6 months. [ID 802]*
Manca d'informació de l'equip de recerca	Experiència científica	Negatiu	*The researchers did not follow instructions and provide information on their team. Instead, they just provided a list of publications. They did describe a thesis, which I presume that members of the team were mentors for these projects but this was not clear. They also provided a list of funding but it was not clear who received which funding. It is impossible to evaluate the researchers' expertise and experience fully with what they provided. [ID 2633]*
Manca d'informació de la proposta	Projecte	Positiu	*I think this is a very interesting and important topic but it is unclear how the researchers are going to disseminate the environmental information to patients but of greater importance, there is no description on how any impact such information might have on asthma related outcomes. Please see my initial review for further methodological issues. [ID 523]*
		Negatiu	*In the description […] I could not find any data about the stage of the tumors (patients) which have been transplanted. This is important to know to understand action or failure of anti-cancer drugs. [ID 1943]*

ID: identificador del projecte

Factors associats amb un canvi substancial de l'avaluació

En el model multinomial ajustat, cinc factors es varen associar estadísticament amb algun tipus de canvi ('millora' o 'empitjorament') entre la valoració qualitativa primera i segona. Una avaluació positiva de l'experiència de l'IP/equip de recerca va mostrar l'associació més forta amb un canvi positiu (RRR=2,63; p<0,0001) i va ser menys probable de tenir un canvi negatiu (RRR=0,25; p<0,0001). Les edicions antigues (2001-2007) junt amb l'edició del 2009 també es van considerar un factor associat a un canvi postitiu, i només es varen associar canvis negatius en les edicions de 2002 i 2004. En comparació amb els avaluadors europeus, un avaluador de Nordamèrica és menys probable que faci un canvi positiu (RRR=0,62; p=0,01). En canvi, quan es demana un pressupost de més de 300.000€, hi ha menys probabilitat de tenir un canvi negatiu (RRR=0.49; p=0,01). Els IP dones també va ser un factor estadísticament significatiu relacionat amb un canvi negatiu (RRR=1,42; p=0,01) en comparació amb els IP homes. Altres factors com l'àrea de recerca de la proposta, l'edat de l'IP, el sexe de l'avaluador o l'índex H de l'avaluador no van resultar estadísticament associats a cap canvi positiu o negatiu (Taula 6).

Taula 6. Factors explicatius que preveuen canvis entre la primera avaluació (projecte anonimitzat) i la segona avaluació (amb informació de l'IP/equip de recerca)

	N Avaluacions (% total)	N Canvi positiu (% categoria)	N Canvi negatiu (% categoria)	Model ajustat RRR (IC95%) Canvi positiu	Canvi negatiu
Any d'edició					
2001	214 (4,28)	41 (19,16)	15 (7,01)	3,83 (1,91-7,68)*	1,45 (0,58-3,58)
2002	160 (3,20)	21 (13,13)	18 (11,25)	2,55 (1,20-5,40)*	2,47 (1,04-5,86)*
2003	231 (4,62)	33 (14,29)	24 (10,39)	3,03 (1,53-6,03)*	2,16 (0,96-4,88)
2004	586 (11,73)	78 (13,31)	61 (10,41)	2,35 (1,28-4,32)*	2,42 (1,18-4,99)*
2005	451 (9,03)	96 (21,29)	29 (6,43)	4,02 (2,19-7,37)*	1,20 (0,55-2,61)
2006	185 (3,70)	22 (11,89)	10 (5,41)	2,33 (1,13-4,80)*	0,88 (0,34-2,25)
2007	356 (7,13)	56 (15,73)	22 (6,18)	2,78 (1,46-5,32)*	1,48 8(0,64-3,41)
2008	339 (6,79)	29 (8,55)	23 (6,78)	1,03 (0,47-2,24)	1,07 (0,42-2,68)
2009	540 (10,81)	74 (13,70)	39 (7,22)	2,36 (1,29-4,34)*	1,94 (0,93-4,07)
2010	222 (4,44)	14 (6,31)	11 (4,95)	0,98 (0,45-2,17)	0,73 (0,28-1,88)
2011	457 (9,15)	29 (6,35)	19 (4,16)	0,94 (0,48-1,85)	0,80 (0,35-1,79)
2012	703 (14,07)	61 (8,68)	28 (3,98)	1,44 (0,78-2,67)	0,79 (0,37-1,70)
2013	326 (6,53)	25 (7,67)	19 (5,83)	1,20 (0,60-2,39)	1,19 (0,53-2,72)
2014	225 (4,50)	15 (6,67)	10 (4,44)	Ref.	Ref.
Àrea de recerca del projecte					
Bàsica	2.387 (47,79)	309 (12,95)	159 (6,66)	0,99 (0,62-1,57)	1,09 (0,61-1,98)
Clínica	1.354 (27,11)	143 (10,56)	95 (7,02)	1,03 (0,80-1,32)	1,21 (0,84-1,73)
Epidemiològica	306 (6,13)	36 (11,76)	25 (8,17)	0,88 (0,65-1,18)	1,01 (0,68-1,51)
Combinacions	948 (18,98)	106 (11,18)	49 (5,17)	Ref.	Ref.
Quantitat de subvenció sol·licitada					
<100.000€	471 (9,44)	45 (9,55)	46 (9,77)	Ref.	Ref.
100.000-199.999€	2.503 (50,15)	294 (11,75)	189 (7,55)	1,22 (0,84-1,76)	0,98 (0,66-1,45)
200.000-299.999€	929 (18,61)	104 (11,19)	53 (5,71)	1,28 (0,84-1,95)	0,73 (0,45-1,19)
≥300.000€	1.088 (21,80)	150 (13,79)	40 (3,68)	1,49 (0,99-2,25)	0,49 (0,29-0,82)*
Sexe de l'IP					
Dona	1.661 (33,25)	190 (11,44)	126 (7,59)	0,99 (0,81-1,21)	1,42 (1,10-1,83)*
Home	3.334 (66,75)	404 (12,12)	202 (6,06)	Ref.	Ref.
Edat de l'IP					
≤40	799 (16,76)	103 (12,89)	58 (7,26)	1,04 (0,81-1,33)	1,10 (0,80-1,51)
>40	3.969 (83,24)	459 (11,56)	253 (6,37)	Ref.	Ref.
Valoració de l'experiència adecuada de l'equip investigador					
Adecuada	4.310 (87,57)	562 (13,04)	213 (4,94)	2,63 (1,73-4,00)*	0,25 (0,19-0,32)*
No adecuada	612 (12,43)	28 (4,58)	111 (18,14)	Ref.	Ref.
Sexe de l'avaluador					
Dona	933 (18,84)	109 (11,68)	59 (6,32)	1,02 (0,80-1,29)	0,88 (0,64-1,22)
Home	4.018 (81,16)	484 (12,05)	268 (6,67)	Ref.	Ref.
Regió mundial de l'avaluador					
Europa	2.838 (56,82)	322 (11,35)	208 (7,33)	Ref.	Ref.
Nord-Amèrica	1.697 (33,97)	228 (13,44)	101 (5,95)	0,62 (0,43-0,89)*	0,65 (0,37-1,15)
Altres	460 (9,21)	44 (9,57)	19 (4,13)	0,84 (0,69-1,03)	1,27 (0,96-1,69)
H-index de l'avaluador (experiència)					
≤15	2.279 (46,95)	292 (12,81)	164 (7,20)	1,08 (0,88-1,32)	1,09 (0,83-1,43)
>15	2.575 (53,05)	287 (11,15)	154 (5,59)	Ref.	Ref.

N: nombre; IP: investigador principal; RRR, Reducció Relativa del Risc; IC, interval de confiança; ref.: variable de referència per al càlcul de la RRR. *Resultats estadísticament significatius. Nagelkerke R^2 = 0,102

VALIDESA D'UNA EINA D'AVALUACIÓ EX-POST DE LA RECERCA

Descriptiu

La taxa de resposta del qüestionari enviat als IPs va ser del 42,8% (n = 68). Els participants varen trigar, de mitjana, 26 minuts en respondre el qüestionari. La taula 7 mostra les característiques descriptives de la mostra, comparant els IPs que van participar amb els que no van participar. L'anàlisi d'aquesta comparació va demostrar diferències estadísticament significatives en l'àrea de coneixement (p = 0,014) i en els grups d'edat (p = 0,047). Els IPs en arts i humanitats i els IPs majors de 50 anys van ser més freqüents entre els no participants. La proporció de dones no va diferir significativament entre els participants i els no participants (p = 0,083).

Taula 7. Característiques descriptives dels participants i dels no participants

		Participants n=68 (42,8%)	No participants n=91 (57,2%)
Disciplina*	Ciències socials	42 (61,8)	47 (51,6)
	Tecnologies de la informació i la comunicació	14 (20,6)	16 (17,6)
	Ciències de la salut	7 (10,3)	7 (4,4)
	Arts i humanitats	5 (7,4)	24 (26,4)
Tema orincipal de la recerca[a]	Educació	25 (36,8)	-
	Internet, tecnologies digitals i mitjans de comunicació	20 (29,4)	-
	Computació i intel·ligència artificial e	15 (22,1)	-
	Estils de vida sostenibles i salut	14 (20,6)	-
	Art, cultura i identitat	11 (16,2)	-
	Societat, acció social i medi ambient	10 (14,7)	-
	Governança i moviments socials	7 (10,3)	-
	Gestió, sistemes i serveis en informació i comunicacions	4 (5,9)	-
	Globalització, pluralisme jurídic i drets humans	4 (5,9)	-
	Llengua, literatura i cognició	4 (5,9)	-
	Turisme	2 (2,9)	-
	Altres	8 (11,8)	-
Impuls original per fer la recerca[a]	La curiositat científica	32 (47,1)	-
	Una necessitat d'omplir certes llacunes en el coneixement	38 (55,9)	-
	L'orientació cap a la pràctica	39 (57,4)	-
	La pròpia experiència com a professional	23 (33,8)	-
	Encàrrec de tercers	2 (2,9)	-
Temps transcorregut des de l'inici de la recerca (anys)	< 4	17 (24,5)	-
	4-9	35 (50,7)	-
	> 9	16 (23,9)	-
	Desconegut	1 (1,4)	-
Interacció am els usuaris finals de la recerca[a]	Abans de començar la recerca	24 (35,3)	-
	Durant la recerca	48 (70,6)	-
	Un cop finalitzada la recerca	41 (60,3)	-
	Cap interacció	7 (10,3)	-
Sexe de l'IP	Dona	34 (50,0)	33 (36,3)
	Home	34 (50,0)	58 (63,7)
Edat de l'IP (anys)*	< 30	3 (4,4)	2 (2,2)
	31-40	17 (25,0)	12 (13,2)
	41-50	38 (55,9)	45 (49,5)
	>50	10 (14,7)	28 (30,8)
	Desconegut	-	4 (4,4)
Experiència de l'IP en recerca (anys)	≤ 5	5 (7,3)	-
	6-10	13 (19,1)	-
	>10	50 (73,5)	-

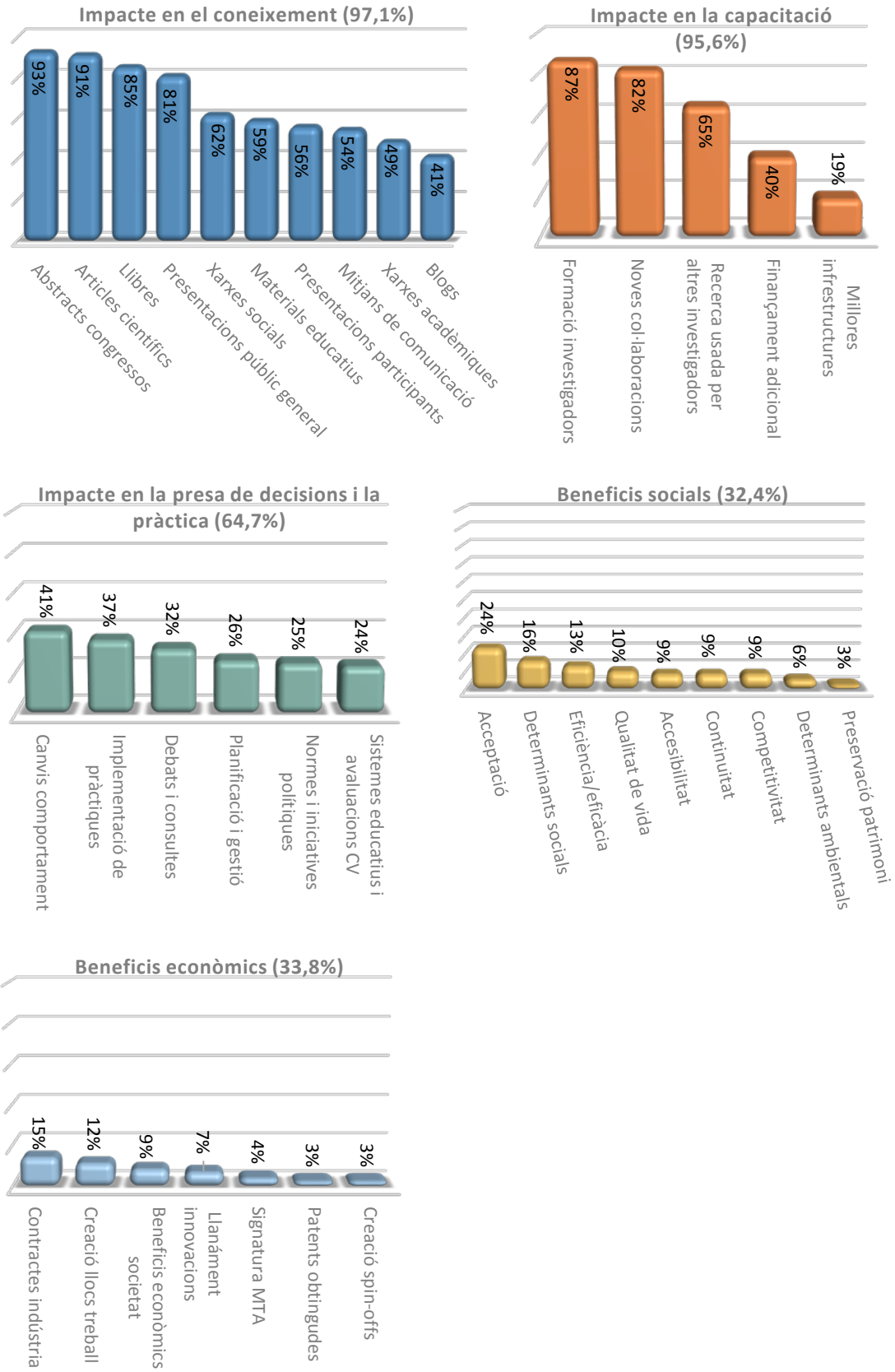
[a] Les respostes poden incloure més d'una categoria; IP: investigador principal; *p≤ 0,05;

Impactes percebuts i factors associats

Els impactes més observats varen ser l'*impacte en el coneixement,* percebut en un 97,1% dels projectes, i l'*impacte en la capacitació*, trobat en un 95,6%. Amb percentatges inferiors li segueix l'*impacte en la presa de decisions i la pràctica*, en un 64,7%, els *beneficis econòmics* (33,8%) i per últim, els *beneficis socials* (32,4%), tot i que en aquests dos últims casos, en només un 11,8% dels projectes, l'impacte es va basar en una avaluació formal. Es va estimar que l'aportació dels projectes als diferents nivells d'impacte era significativa en el coneixement (entre el 70% i el 100%), moderada en la capacitació (entre el 40% i el 60%), i limitada (1-30%) per informar polítiques i pràctiques, beneficis socials i beneficis econòmics. Cap dels IPs va descriure altres impactes addicionals.

El gràfic 7 mostra les diferents categories d'impacte i la distribució de les subcategories d'impacte. La mida de les barres indica el percentatge de projectes en què es va produir aquest impacte específic, d'acord amb les respostes dels IPs.

Gràfic 7. Impactes percebuts, segons els diferents nivells (n=68)



L'impuls original del projecte es va correlacionar estadísticament amb diferents nivells d'impacte: quan els projectes estaven destinats a omplir certes llacunes de coneixement, es va observar un major *impacte en el coneixement* (p = 0,01) i en la *capacitació* (p = 0,03), mentre que quan els projectes estaven orientats a una aplicació pràctica, es va observar més impacte en la *presa de decisions i la pràctica* (p = 0,05) i en els *beneficis socials* (p = 0,01). En general, els projectes que interactuaven amb els usuaris finals, en qualsevol moment del procés de recerca, van tenir més *impacte en el coneixement* (p = 0,01), en la *capacitació* (p = 0,03) i en els *beneficis socials* (p = 0,05). Majors *impactes en el coneixement* i en la *capacitació* es van correlacionar amb els IPs majors de 40 anys o que tenien més de tres anys d'experiència en recerca (p ≤ 0,01). També, els projectes que havien començat feia 4 anys o més es va correlacionar amb més *impacte en el coneixement* (p= 0,04) i en la *capacitació* (p= 0,04). En canvi, el sexe de l'IP no es va correlacionar amb cap nivell d'impacte. En la següent taula es pot veure la magnitud de les diferències entre els projectes que no varen tenir impacte i els projectes que sí, per a cadascun dels nivells (Taula 8). S'ha de tenir en compte el baix nombre de casos amb impacte en el *coneixement* i en la *capacitació*.

Taula 8. Percentatge de projectes amb i sense impacte, segons els diferents factors

	IMPACTE EN EL CONEIXEMENT			IMPACTE EN LA CAPACITACIÓ			IMPACTE EN PRESA DE DECISIONS I PRÀCTICA			BENEFICIS SOCIALS			BENEFICIS ECONÒMICS		
	Sense impacte (n=2)	Amb impacte (n=66)	Valor de p	Sense impacte (n=3)	Amb impacte (n=65)	Valor de p	Sense impacte (n=24)	Amb impacte (n=44)	Valor de p	Sense impacte (n=46)	Amb impacte (n=22)	Valor de p	Sense impacte (n=45)	Amb impacte (n=23)	Valor de p
Impuls original per fer la recerca															
Curiositat científica	50,0	10,6	0,01*	66,7	9,2	0,03*	25,0	4,5	0,06	17,4	0	0,01*	13,3	8,7	0,51
Omplir llacunes en el coneixement	0,0	22,7		0,0	23,1		25,0	20,5		28,3	9,1		24,4	17,4	
Orientació cap a la pràctica	50,0	57,6		33,3	58,5		41,7	65,9		43,5	86,4		51,1	69,6	
Pròpia experiència com a professional	0,0	9,1		0,0	9,2		8,3	9,1		10,9	4,5		11,1	4,3	
Temps transcorregut des de l'inici de la recerca (anys)															
< 4 anys	100,0	21,2	0,04*	100,0	2,3	0,04*	34,8	18,2	0,18	26,1	18,2	0,43	24,4	21,7	0,66
4-9 anys	0,0	53,0		0,0	20,5		52,2	52,3		45,7	63,6		53,3	47,8	
> 9 anys	0,0	24,2		0,0	77,3		13,0	29		26,1	18,2		20,0	30,4	
Interacció am bels usuaris finals de la recerca															
Abans de començar	100,0	10,6	0,01*	66,7	10,8	0,03*	12,5	13,6	0,70	13,0	13,6	0,05*	15,6	8,7	0,27
Durant la recerca	0,0	43,9		0,0	44,6		50,0	38,6		45,7	36,4		35,6	56,5	
Un cop finalitzada	0,0	12,1		33,3	10,8		12,5	11,4		13,0	9,1		15,6	4,3	
En tot el procés	0,0	22,7		0,0	23,1		12,5	27,3		13,0	40,9		20,0	26,1	
Cap interacció	0,0	10,6		0,0	10,8		12,5	9,1		15,2	0,0		13,3	4,3	
Sexe de l'IP															
Dona	50,0	50,0	1,00	66,7	49,2	0,56	54,2	47,7	0,61	54,3	40,9	0,30	44,4	60,9	0,20
Home	50,0	50,0		33,3	50,8		45,8	52,3		45,7	59,1		55,6	39,1	

Taula 8. Percentatge de projectes amb i sense impacte, segons els diferents factors (continuació)

	IMPACTE EN EL CONEIXEMENT			IMPACTE EN LA CAPACITACIÓ			IMPACTE EN PRESA DE DECISIONS I PRÀCTICA			BENEFICIS SOCIALS			BENEFICIS ECONÒMICS		
	Sense impacte (n=2)	Amb impacte (n=66)	Valor de p	Sense impacte (n=3)	Amb impacte (n=65)	Valor de p	Sense impacte (n=24)	Amb impacte (n=44)	Valor de p	Sense impacte (n=46)	Amb impacte (n=22)	Valor de p	Sense impacte (n=45)	Amb impacte (n=23)	Valor de p
Edat de l'IP (anys)															
< 30	50,0	3,0	0,01*	66,7	1,5	<0,01*	8,3	2,3	0,06	4,3	4,5	0,46	6,7	0,0	0,31
31-40	50,0	24,2		33,3	24,6		25,0	25,0		30,4	13,6		28,9	17,4	
41-50	0,0	57,6		0,0	58,5		66,7	50,0		50,0	68,2		48,9	69,6	
>50	0,0	15,2		0,0	15,4		0,0	22,7		15,2	13,6		15,6	13,0	
Experiència de l'IP en recerca (anys)															
≤5	50,0	6,1	<0,01*	66,7	4,6	<0,01*	16,7	2,3	0,15	10,9	0	0,32	8,9	4,3	0,56
6-10	50,0	18,2		33,3	18,5		16,7	20,5		15,2	27,3		22,2	4,3	
>10	0,0	75,8		0,0	76,9		66,7	77,3		73,9	72,7		68,9	13,0	

*p≤ 0,05

Validesa de l'instrument

Per a mesurar la consistència interna de les preguntes del qüestionari es va utilitzar l'alfa de Cronbach. Aquesta va ser satisfactòria (α = 0,89) en el conjunt de les preguntes. La taula 9, mostra el seu valor per a cadascun dels nivells d'impacte. La consistència interna va ser satisfactòria en tots els nivells d'impacte, a excepció dels *beneficis econòmics* Tanmateix, l'eliminació de qualsevol de les preguntes d'aquest nivell d'impacte hauria donat com a resultat una alfa de Cronbach igual o inferior.

Taula 9. Consistència interna per a cada nivell d'impacte

Nivell d'impacte	Alfa de Cronbach
Impacte en el coneixement	0,74
Impacte en la capacitació	0,74
Impacte en la presa de decisions i la pràctica	0,82
Beneficis socials	0,89
Beneficis econòmics	0,47
Totes les preguntes	**0,89**

Per a la validesa de contingut es va rebre resposta de 13 dels 17 experts contactats (76,5%), els quals varen avaluar mitjançant un formulari si les preguntes del qüestionari d'impacte eren rellevants per a l'objectiu de l'estudi. Set dels experts eren de ciències socials (54%), quatre de ciències de la salut (31%) i dos de tecnologies de la informació i la comunicació (15%). El 39% (n=5) eren dones, i tots tenien una llarga experiència com a investigadors o gestors de recerca. A més, varen afegir un nou ítem al qüestionari, el relacionat amb 'noves xarxes acadèmiques'. Els experts van puntuar els 45 ítems del qüestionari d'impacte segons la seva rellevància i el 76% de les qualificacions (n = 34) varen obtenir un índex igual o superior a 0,5, considerat el nivell acceptable en aquest estudi. En la taula 10 es mostra l'índex de cada ítem, tant del conjunt d'experts, com segmentats per la seva expertesa.

Taula 10. Índex de validesa de contingut segons els experts

Ítem	Nombre de respostes	Tots els experts (n=13)	Experts ciències socials (n=7)	Experts ciències salut (n=4)	Experts TIC (n=2)
Impacte en el coneixement					
Resums de congressos	13	0,77*	0,71*	0,75*	1,00*
Articles científics	13	0,85*	0,71*	1,00*	1,00*
Llibres o capitols de llibres	13	0,69*	0,71*	0,50*	1,00*
Materials educatius	13	0,62*	0,57*	0,50*	1,00*
Presentacions al públic general	13	0,62*	0,43	1,00*	0,50*
Presentació als participants de la recerca	13	0,69*	0,57*	0,75*	1,00*
Mencions en els mitjans de comunicació	13	0,85*	0,71*	1,00*	1,00*
Xarxes socials	13	0,62*	0,43	1,00*	0,50*
Blogs influents	12	0,67*	0,57*	1,00*	0,50*
Concerts, enregistraments o sales de música	9	0,11	0,25	0,00	0,00
Impacte en la capacitació					
Formació d'estudiants de doctorat	13	0,92*	0,86*	1,00*	1,00*
Formació d'estudiants de màster	13	0,69*	0,57*	0,75*	1,00*
Formació d'estudiants de grau	13	0,31	0,43	0,00	0,50*
Noves col·laboracions a nivell nacional	11	0,91*	0,83*	1,00*	1,00*
Noves col·laboracions a nivell internacional	12	0,92*	0,83*	1,00*	1,00*
Noves xarxes acadèmiques	8	0,75*	0,75*	0,50*	1,00*
Finançament addicional per a nous projectes de recerca	13	0,77*	0,57*	1,00*	1,00*
Finançament addicional pel grup de recerca	13	0,92*	0,86*	1,00*	1,00*
Recerca o mètodes utilitzats per altres investigadors	9	0,78*	0,67*	0,75*	1,00*
Millores en les infraestructures	13	0,62*	0,57*	0,50*	1,00*
Impacte en la presa de decisions I la pràctica					
Informant en processos de discussions, assessories o debats per a la presa de decisions	13	0,62*	0,71*	0,50*	0,50*
Informant en la formulació de normes, iniciatives polítiques o recomanacions per a reguladors	12	0,58*	0,67*	0,50*	0,50*
Contribució en el disseny, planificació i gestió de serveis i prioritats	12	0,50*	0,50*	0,50*	0,50*
En la implementació, adopció o producció de pràctiques fora i dintre del món professional	12	0,58*	0,50*	0,50*	1,00*
Influint el comportament dels professionals o d'altres	13	0,54*	0,57*	0,50*	0,50*
Influint els sistemes d'educació i les avaluacions curriculars	10	0,40	0,50*	0,25	0,50*

Taula 10. Índex de validesa de contingut segons els experts (continuació)

Ítem	Nombre de respostes	Tots els experts (n=13)	Experts ciències socials (n=7)	Experts ciències salut (n=4)		Experts TIC (n=2)
Beneficis socials						
En la salut	11	0,27	0,20	0,50*		0,00
En la qualitat de vida	11	0,55*	0,60*	0,50*		0,50*
En determinants socials I culturals	11	0,64*	0,80*	0,50*		0,50*
En determinants ambientals	11	0,36	0,40	0,50*		0,00
Acceptació	11	0,73*	0,80*	0,75*		0,50*
Accessibilitat	11	0,55*	0,60*	0,50*		0,50*
Continuïtat	9	0,56*	0,67*	0,50*		0,50*
Efectivitat i eficiència	10	0,50*	0,50*	0,50*		0,50*
Seguretat	10	0,20	0,25	0,25		0,00
Benestar I beneficis socials	11	0,64*	0,80*	0,50*		0,50*
Preservació del patrimoni	10	0,20	0,25	0,25		0,00
Competitivitat i desenvolupament d'estímuls	10	0,50*	0,50*	0,25		1,00*
Beneficis econòmics						
Obtenint alguna patent	13	0,23	0,00	0,50*		0,50*
Rebent ingressos per royalties, contractes amb la industria i/o prestació de serveis	13	0,23	0,14	0,50*		0,00
En la creació d'empreses spin-offs i/o start-ups	13	0,31	0,14	0,50*		0,50*
En la creació d'acord de Material Transfer Agreements (MTS)	13	0,23	0,00	0,50*		0,50*
Portant al mercat innovacions, productes o dispositius produïts pel sector privat	9	0,56*	0,33	0,50*		1,00*
Creant nous llocs de treball	13	0,62*	0,57*	0,75*		0,50*
Beneficis econòmics cap a la societat	13	0,54*	0,57*	0,50*		0,50*

** elements amb una puntuació superior o igual a 0,5; TIC: tecnologies de la informació i la comunicació*

Quasi tots els ítems del nivell l'*impacte en el coneixement*, menys un, es van considerar acceptables (índex d'opinió d'experts ≥ 0,5). En l'*impacte en la capacitació* el 89% es van considerar acceptables, així com un 83% dels ítems en l'*impacte en la presa de decisions i la pràctica* i un 63% dels ítems de *beneficis socials*. En canvi, només el 43% dels ítems (tres de cada set) del nivell de *beneficis econòmics* van aconseguir una qualificació acceptable. Alguns dels ítems varen ser més rellevants segons experts de disciplines específiques, que no pas en el conjunt d'experts. Per exemple, l'ítem relacionat amb la salut van ser considerats acceptables pels experts en salut junt amb la majoria d'ítems en beneficis socials, tot i que aquests últims, els experts en ciències socials encara els van valorar més favorablement; la formació d'alumnes de grau va ser considerada acceptable pels experts en TIC; la influència en els sistemes educatius

i en les avaluacions curriculars, va ser qualificada acceptable pels experts en ciències socials i en TIC; i els ítems relacionats amb la comercialització de productes van ser considerats rellevants pels experts en salut i els de TIC (Taula 10).

Per últim, en relació amb la validesa discriminant, tres dels ítems van tenir efecte sòl, ja que cap IP els va informar com impacte dels seus projectes ('presentacions en concerts, enregistraments o sales de música', 'millora de la seguretat' i 'millora de la salut'). La validesa del contingut d'aquests tres ítems, segons el conjunt d'experts, també va ser deficient. No es va detectar cap efecte sostre.

Capítol 4: Discussió

Aquesta tesi ha desenvolupat diverses estratègies per mesurar la validesa de dos mètodes d'avaluació de la recerca en diferents fases del procés d'investigació. El primer, validant un procés d'avaluació ex-ante relacionat amb una de les fases de revisió per consemblants de projectes de recerca, i el segon, validant una eina d'avaluació ex-post dissenyada per mesurar els impactes de la recerca percebuts pels IPs. A continuació es presenta una discussió dels principals resultats, a partir de les quatre preguntes de recerca de l'estudi.

PREGUNTA DE RECERCA 1 (EX-ANTE: VALIDESA)

Pregunta de recerca 1. *Quin és l'efecte d'anonimitzar la identitat dels investigadors i les seves institucions en la primera etapa d'avaluació d'un procés de revisió per parells de propostes de recerca, sobre el canvi en l'avaluació de l'avaluador un cop aquest coneix, en una segona etapa, el nom de l'investigados principal/grup de recerca, la seva experiència i la institució a la qual pertany?*

L'avaluació dels projectes anonimitzant la identitat dels investigadors i les seves institucions en una primera fase, i la presentació de la identitat d'aquests en una segona fase, fa que els avaluadors canvÏin de mitjana la seva avaluació en un 18,5% de les avaluacions (entre un 10% i un 28% segons l'anualitat). El fet que en gairebé el 19% de les avaluacions es canvii la seva valoració un cop coneguda la identitat dels investigadors i les seves institucions, obre, d'una banda, una nova discussió sobre el temps i el cost de l'anonimització. Aquest percentatge és suficient per continuar insistint i assegurant l'anonimat dels investigadors durant la primera fase d'avaluació de les propostes?

En el disseny del procés d'avaluació estudiat es fan grans esforços per garantir l'anonimització de les propostes en la primera fase. Totes les propostes es verifiquen una per una per tal d'eliminar qualsevol dada que identifiqui als investigadors o a les seves identitats. Tot i que la majoria dels avaluadors manté la valoració de la proposta una vegada coneguda la identitat dels investigadors, el fet que en quasi un 19% de les avaluacions aquesta es modifiqui sembla un

percentatge suficient per continuar insistint i garantint l'anonimització dels investigadors en una primera fase del procés d'avaluació.

Si aquest percentatge per si mateix és suficient o no pot semblar difícil de jutjar degut a la manca de comparadors amb altres estudis. Efectivament, no és gaire freqüent entre les agències finançadores de recerca, incloure una primera fase d'avaluació de propostes anonimitzades, segons ens diu una enquesta de la *European Science Foundation*,[50] però encara és menys habitual, avaluar la seva validesa. Per tant, és necessari tenir en compte d'altres aspectes, més que no pas, el percentatge en si mateix.

Un dels principals problemes de qualsevol sistema de finançament científic és l'èmfasi relatiu que es posa en els individus enfront de les idees. Són molts els retractors de ficar el pes en el finançament d'idees.[51] El motiu principal seria perquè es considera que és difícil que un únic projecte doni lloc a un avenç innovador. Més aviat és necessari que un equip recopili i construeixi la seva recerca al llarg del temps i, per tant, el procés d'avaluació no necessitaria avaluar un projecte de manera aïllada.[52] Naturalment, posar el pes en el talent de les persones o en les propostes dependrà dels objectius de la convocatòria. En el nostre cas, en canvi, el que es vol reforçar és justament les idees, indiferentment de si els investigadors aporten més o menys quantitat a les seves institucions (o com a mínim, no és el factor primordial), per tal d'enfortir la rigorositat i ajudar a disminuir un potencial elitisme.

Per tant, hem aconseguit validar el contingut del procés? És a dir, hem obtingut el resultat que volíem, ficant el pes de l'avaluació en les idees i no pas en les persones? Com han demostrat altres estudis,[53] els resultats d'aquesta tesi semblen reafirmar la importància d'avaluar les característiques d'una proposta (qualitat, originalitat, metodologia, innovació, etc.), ja que les valoracions es van mantenir sense canvis en la majoria dels casos. No obstant això, que 'només' en el 19% de les avaluacions es canviï la valoració no s'ha d'interpretar com una falta de la utilitat de l'anonimització. Fàcilment això es podria explicar pel fet que les propostes de bona qualitat venen escrites per equips amb experiència en el tema, és a dir, un equip potent en un tema, fàcilment escriuria una proposta potent. Fer el canvi, per tant, implica més dubtes en la valoració. Això queda reflectit en la proporció de canvis, en funció de quina és la classificació de la primera avaluació (anonimitzada). Un 25,2% venen d'avaluacions classificades com dubtoses, un 19,4% de finançables amb reserves, un 19,1% de no finançables i un 9,9% de finançables. Aquesta teoria, també vindria reforçada pel fet que de les propostes on es canvia la valoració,

només el 14% s'acaben finançant. Per tant, els resultats d'aquesta tesi semblen suggerir que els canvis són més habituals en propostes pobres que en bones propostes.

No ens ha d'estranyar que els principals motius per realitzar el canvi, segons els comentaris dels avaluadors, és l'experiència de l'IP/equip de recerca, ja que aquesta és la nova informació que reben els avaluadors en la segona fase. Una avaluació positiva o negativa de les habilitats i l'experiència de l'equip investigador o de l'IP, la col·laboració multiinstitucional o la composició i la naturalesa multidisciplinària de l'equip són els principals factors esmentats per canviar l'avaluació. Per tant, l'experiència dels investigadors i les seves institucions té un paper important a l'hora de valorar la viabilitat d'una proposta, és a dir, s'ha de valorar en una segona fase, donant un pes inferior, però no obviar-ho, tal i com afirmen en els resultats d'un estudi de l'agència Nord-americana *National Science Foundation* (NSF),[54] on es considerava que el fet de no conèixer la identitat dels sol·licitants podia comprometre la integritat d'una proposta.

PREGUNTA DE RECERCA 2 (EX-ANTE: FACTORS INFLUENTS)

Pregunta de recerca 2. *Quins* per rebre finançament en les seves investigacions *factors externs afecten el procés de presa de decisions, independentment de la qualitat de la proposta que s'avalua?*

L'objectiu d'anonimitzar les propostes en una primera fase no és canviar la qualitat de la revisió, sinó afavorir una revisió objectiva i justa. Per tant, és important d'objectivar quins són els factors associats al fet de canviar la valoració una vegada es coneix l'equip i les institucions de la proposta. Els principals factors associats al canvi s'han relacionat amb l'IP/equip de recerca, amb els avaluadors i amb el projecte. Els biaixos més comuns en les avaluacions es relacionen amb les característiques pròpies dels investigadors com el sexe, l'edat, el grup minoritari i, especialment, l'efecte conegut com a efecte Mateu, relacionat amb el prestigi, la reputació o el reconeixement d'investigadors o institucions.[55] En el nostre cas, no ens ha d'estranyar que la valoració de l'experiència estigui altament relacionada amb el canvi donat que és la informació addicional que reben els avaluadors en la segona fase d'avaluació. La valoració sobre la informació de l'experiència prèvia com el finançament aconseguit o les publicacions liderades efectivament pot reforçar les dinàmiques relacionades ambl'efecte Mateu on l'èxit del passat afavoreix l'èxit del futur,[55] la qual cosa provoca alhora noves diferències que es van augmentant amb el temps,[56] és a dir, aconsegueixen una avantatge acumulativa. Tot i els nostres resultats, l'anonimització de les propostes en una primera fase el que pretén és minimitzar aquest biaix

des d'un primer moment cap a aquests investigadors 'potents' i com a mínim, no ficar tot el pes de l'avaluació a l'experiència de l'equip sinó que aquesta sigui un punt rellevant a l'hora de mesurar si el projecte es pot portar a terme o no, és a dir, la viabilitat de la proposat en funció de si l'equip de recerca té l'experiència suficient per realitzar els objectius de la proposta. Per exemple, en el cas de projectes innovadors, aquest punt pren gran rellevància, ja que implica la necessitat d'un equip de recerca potent darrere del projecte, capaç de portar-lo a terme, i per tant avaluacions de la proposta que poden ser més dubtoses sobre la viabilitat de la proposta, poden clarificar-se veient l'experiència de l'equip.

En aquesta tesi també es va veure que ser IP dona s'associava a un canvi negatiu en comparació amb cap canvi i respecte als homes. El biaix de gènere continua sent un problema discutit en el procés de revisió per consemblants amb resultats no concloents.[57–60] Entre els estudis que declaren un biaix de gènere hi ha una metanàlisi de 21 estudis que evidencien que les propostes presentades per homes tenen un 7% superior de ser finançdes que les presentades per homes.[60] Altres estudis posteriors també ho correboren.[58,61,62] D'altra banda, altres estudis han conclòs que no existeix el biaix de gènere, per exemple, en una gran metanàlisi[63] i altres estudis posteriors.[64] En el cas d'aquesta tesi, i tot i els resultats associats al canvi negatiu, hem vist que el percentatge d'IP dones que presenten projectes (33,1%) és gairebé exacte al percentatge d'IPs dones que aconsegueixen el finançament (33,1%). Per tant suposem que el que es veu aquí podria no ser una qüestió sobre el peer review sinó un aspecte de la pròpia ciència. La revista The Lancet[65] va definir aquesta qüestió com 'el cicle viciós'. Segons la descripció, les dones tenen més dificultats per rebre finançament en les seves investigacions, impossibilitant que accedeixin a llocs més rellevants en la posició d'autoria dels articles, reben menys invitacions a congressos i conferències, i com a conseqüència, un currículum inferior que, implica menors puntuacions en el finançament, com en aquesta tesi. En qualsevol cas, el disseny del nostre estudi impedeix provar experimentalment el paper del sexe, i l'associació trobada, no implica causalitat.

Pel que fa a les característiques de l'avaluador, en aquesta tesi els avaluadors nord-americans tenen menys risc de fer un canvi de millora, en comparació amb els avaluadors europeus. Això es recolza en la literatura, ja que segons un estudi del Consell de Recerca d'Austràlia,[57] els revisors nord-americans donaven puntuacions més altes que les d'altres països, com ara els d'Europa. Per últim, pels factors relacionats amb els projectes, podrien tenir a veure amb el que és la gestió del mateix procés d'avaluació. D'una banda, les diferències entre les diferents convocatòries podrien estar relacionades en què en convocatòries d'edicions antigues, es van realitzar més modificacions per millorar el procés, les quals podrien tenir a veure amb aquestes

diferències. Per exemple, algunes de les modificacions introduïdes varen implicar que només l'IP presentés tot el currículum, mentre que l'equip de recerca presenta una breu descripció, des del 2005, o l'homogeneïtzació de la presentació de projectes coordinats des de 2007. És a dir, en les edicions més recents, el procés de revisió per consemblants ha estat més estandarditzat i només s'han aplicat modificacions menors. D'altra banda, les diferències també podrien estar relacionades no tant amb l'anualitat de la proposta, sinó amb si la recerca que es fa en el camp de la convocatòria és més o menys competitiva. És dir, en anualitats amb taxes d'èxit més baixes, els avaluadors poden dubtar més sobre, per exemple, projectes més arriscats. Pel que fa a les diferències en funció del pressupost sol·licitat es poden explicar si tenim en compte que els imports més baixos estan relacionats amb propostes d'un sol grup. Això significa que, en general, si s'inclouen un nombre menor d'equips de recerca/IP en una proposta (i per tant, es demana un pressupost inferior), serà més difícil compensar l'especialització i la inclusió de diferents disciplines.

Aquesta tesi també mostra que el canvi en la valoració de la segona fase no es veu afectat per altres característiques que, segons la literatura, podrien comportar algun biaix, com l'àrea de recerca del projecte, el sexe del revisor, l'experiència del revisor (mesurada amb l'índex H), o l'edat de l'IP. Pel que fa aquest últim, tot i que hi ha poca i contradictòria evidència sobre si els processos de revisió estan esbiaixats per l'edat,[66,67] no deixa de ser un aspecte que preocupa principalment per l'oportunitat que s'acaba donant als investigadors novells, que poden veure's en desavantatge per no tenir resultats preliminars o un gran llistat de publicacions.

Els factors aquí trobats afavoreixen l'anonimització de l'equip de recerca i les seves institucions en una primera fase de l'avaluació per tal de millorar la fiabilitat en la presa de decisions, la credibilitat i l'equitat del procés d'avaluació. Tot i això, hi ha factors que provenen de les mateixes al·legacions de parcialitat que es troben en el món de la ciència en general. És important, però, tenir clar precisament on es troben aquestes inequitats, per tal de prendre mesures per atenuar aquestes amenaces, corregir-les i millorar la fiabilitat del procés.

PREGUNTA DE RECERCA 3 (EX-POST: VALIDESA)

Pregunta de recerca 3. *El qüestionari d'avaluació ex post dels projectes per mesurar l'impacte social de la recerca és vàlid?*

En aquesta tesi s'han provat les propietats mètriques d'un qüestionari dissenyat per registrar l'impacte de la recerca universitària provinent de diverses disciplines. El qüestionari ha mostrat una bona consistència interna i valideses de contingut i discriminant acceptables en el context estudiat.

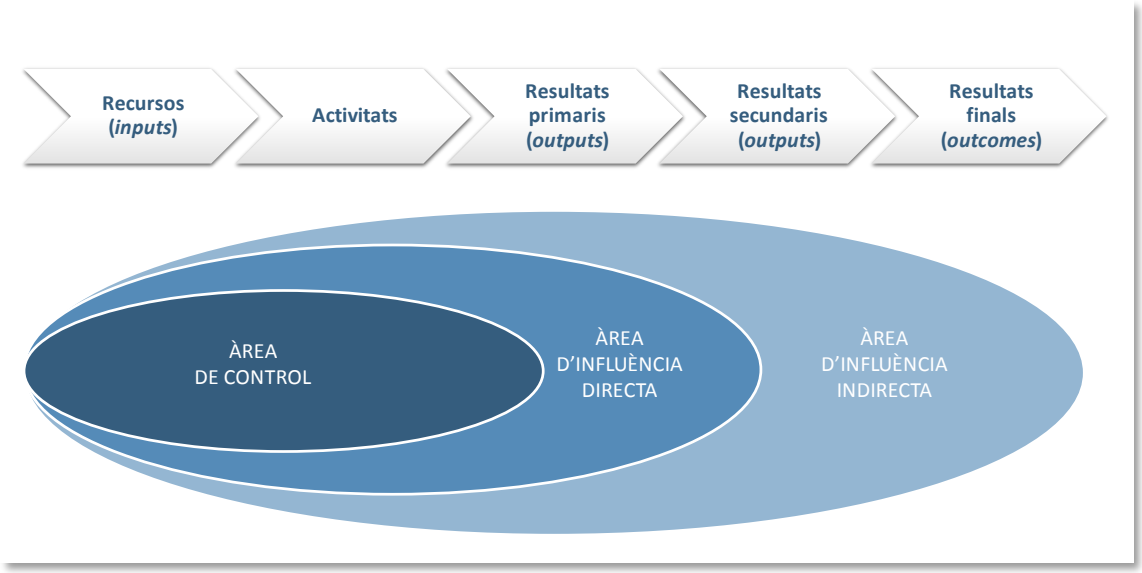
La consistència interna de tots els ítems de l'instrument va resultar excel·lent demostrant la mesura d'un mateix constructe. Tanmateix, atès que el concepte d'impacte inclou una multidimensionalitat, es va avaluar la consistència interna per a cadascun dels cinc nivells d'impacte. Aquesta es va considerar excel·lent en tots els casos excepte en el nivell de *beneficis econòmics*. Aquesta baixa consistència interna podria estar relacionada amb què aquest nivell d'impacte estava format per relativament pocs ítems. Una possibilitat alternativa és que l'anàlisi de la consistència interna ve determinat per l'adequació o rellevància del contingut dels ítems. En aquest sentit, semblaria que aquesta anàlisi pot ser útil per eliminar aquells ítems 'dolents', quan aquesta és baixa, tot i que en el cas de la dimensió de *beneficis econòmics*, s'ha comprovat que l'eliminació de qualsevol dels ítems no afectava la consistència interna del qüestionari.

Hi ha un consens en la literatura científica que la validesa de contingut no és una propietat de l'instrument, sinó de la interpretació de l'instrument i, per tant, una qüestió de judici.[68] En aquesta tesi hem incorporat dues fases per avaluar aquesta validesa. D'una banda, l'avaluació dels experts sobre la rellevància dels ítems del qüestionari va demostrar que alguns dels ítems no eren considerats acceptables. Els experts que van participar han estat útils per comptar amb una mostra d'opinions per a la selecció dels millors indicadors d'impacte, o d'alguna manera, els més comuns i rellevants en el nostre context. Els experts van mostrar que si bé la majoria dels ítems es podrien considerar 'genèrics' a qualsevol disciplina i, per tant, rellevants en totes elles, alguns dels ítems es podrien considerar més propis de determinades disciplines, i per tant, específics en un context determinat. Aquesta qüestió també demostraria el perquè de l'efecte sòl en tres dels ítems.

Hem de considerar, d'una banda, que utilitzar experts per a la selecció dels indicadors del qüestionari el que permet és recomanar un conjunt equilibrat d'indicadors, reduir el nombre de preguntes del qüestionari (i per tant la càrrega dels IPs que contesten) i centrar-se en els impactes d'interès. D'altra banda, la validesa de contingut ha estat més freqüent en els nivells 'més propers' als investigadors, és a dir, els impactes que tradicionalment s'han utilitzat per mesurar la recerca, i aquells en què els investigadors tenen un cert control i influència (ja que els coneixen més i són més conscients d'ells).[69] En altres paraules, la comprensió del concepte

d'impacte en el *coneixement*, en la *capacitació* i en la *presa de decisions informada i la pràctica*, és a dir, els impactes en nivells intermedis (resultats secundaris o *outputs*) mostra una major homogeneïtat que en els impactes més llunyans (resultats finals o *outcomes*) (Gràfic 8).

Gràfic 8. Cercles d'influència



Per tant, què seria més efectiu o vàlid, disminuir el nombre d'indicadors del qüestionari (als més rellevants), o incloure'n de més, encara que aquests siguin menys rellevants, donat que els investigadors coneixen (i són conscients) principalment dels impactes que tenen sota la seva àrea d'influència? Clarament la resposta vindria donada en funció dels objectius de l'avaluació, els quals necessiten una reflexió contínua i determinen la metodologia i els indicadors d'avaluació utilitzats.[20] Hem de considerar que els índexs d'acord, com el de validesa de contingut utilitzat en aquesta tesi, només són una part del procés de valoració de la validesa del contingut de l'instrument, i no necessàriament hauran de ser l'únic motiu per rebutjar o incloure un indicador. En un primer moment, també es va avaluar la validesa de contingut amb l'anàlisi i el mapatge dels impactes dels casos REF2014 seleccionats aleatòriament. Aquesta fase va ajudar no només a millorar la conceptualització d'alguns dels ítems del qüestionari, sobretot, en l'*impacte en la presa de decisions i la pràctica* i en els *beneficis socials,* sinó també, a categoritzar impactes en tots els indicadors del qüestionari. Els impactes de la recerca són diversos i poden ser inesperats i aparèixer en qualsevol àmbit. Limitar el nombre d'indicadors als més rellevants i comuns podria fer que no es contemplin beneficis inesperats de la recerca.[70] A més, el fet que només es produeixi un impacte en un únic projecte no implica que aquest sigui menys rellevant,

sinó que aquest és menys comú. Per tant, la modesta rellevància d'alguns dels indicadors, podria explicar-se per les característiques de la mostra, és a dir, de quina és la recerca que s'ha avaluat.

El fet d'utilitzar un qüestionari estructurat implica una manera sistèmica d'avaluar l'impacte que afavoreix la seva utilitat. La intenció dels enfocaments quantitatius com els qüestionaris és superar els perills de la subjectivitat. És ben clar que els indicadors més a prop de l'àrea de control dels investigadors (com publicacions o tesis realitzades) tenen definicions més estandarditzades i conegudes per a tothom, mentre que calen més esforços per captar impactes més difícils de quantificar (com les influències en les polítiques i la pràctica, o els canvis en el comportament). A l'hora d'escollir els indicadors utilitzats es recomana captar tota la gamma d'impactes i beneficis; permet agregar i desagregar els impactes en dimensions; que la càrrega per als investigadors sigui baixa; poder captar i comparar la informació de manera equitativa entre diferents ajuts o tipus de recerca; i proporcionar suficient temps perquè es produeixi l'impacte.[71] L''evidència de la literatura afirma que és important considerar un enfocament polifacètic en l'avaluació de la recerca.[72]

Finalment, val a dir que un qüestionari d'avaluació de l'impacte de la recerca no és un instrument de mesura típic on es poden avaluar totes les propietats mètriques, ja que el resultat d'aquests tipus d'instruments no es poden comparar amb cap estàndard, al dependre de les característiques del context. És a dir, els impactes de la recerca podrien variar depenent de les característiques dels IPs, del seu coneixement sobre l'impacte, de les particularitats de la recerca, o del moment en que es realitza l'avaluació.

PREGUNTA DE RECERCA 4 (EX-POST: FACTORS INFLUENTS)

Pregunta de recerca 4. *Quines són les característiques dels projectes més influents en els impactes assolits?*

El qüestionari utilitzat ha permès donar valor a la recerca que es fa a la UOC en concepte d'avenç en el coneixement, formació d'investigadors, noves col·laboracions i projectes, a més d'examinar altres impactes més llunyans. Avaluat en el seu conjunt, els projectes de recerca valorats representen un estímul per a la producció de coneixement i el desenvolupament de competències de recerca en individus i equips. Els projectes han produït, per tant, coneixement per informar principalment a altres investigadors i, en menor mesura, a públics no acadèmics. També, la recerca ha contribuït al desenvolupament individual o de l'equip d'una millor capacitació en recerca. Aquest punt també queda explorat en la pregunta sobre la contribució

de la recerca als diferents nivells d'impacte on els IPs valoren més elevadament la contribució que fan els projectes als beneficis per als equips i institucions de recerca més que als beneficis administratius o polítics, socials o econòmics. Això posa de manifest el reconeixement del valor de la recerca i dels seus beneficis en el potencial de nova recerca de la universitat.

En canvi, pel que fa a la resta d'impactes no acadèmics, és a dir, la utilització efectiva dels resultats de la recerca i els beneficis polítics, socials i econòmics dels resultats de la recerca, s'ha contribuït de manera més limitada. Per tant, hi ha una marcada disparitat entre, d'una banda, els beneficis en l'impacte acadèmic i científic (coneixement i capacitació), notablement elevats i uniformes, i la contribució a beneficis polítics, socials i econòmics, variable i desigual al llarg dels projectes. Aquesta disparitat és coherent amb les troballes d'altres estudis on es mostra que els bons assoliments dels resultats en l'àmbit acadèmic, no necessàriament impliquen uns bons assoliments fora de l'àmbit acadèmic.[73,74] Tot i això, la diversitat d'impactes trobats en aquesta tesi és sovint trobat en altres estudis que també tenen l'objectiu de retiment de comptes.[11,14]

Que els impactes en la presa de decisions, socials i econòmics només s'hagin donat en un grup minoritari de projectes, no ens ha de sorprendre, ja que aquests, no apareixen de manera regular i homogènia en els projectes, sinó que el seu recorregut dependrà de l'àmbit i les circumstàncies afavoridores que es produeixin. El temps que ha de transcorre entre que la recerca es produeix fins que s'aconsegueix un impacte específic és molt variable i difícil de determinar, ja que dependrà del context on es desenvolupi. Tot i això, que el temps des que va començar el projecte sigui un factor afavoridor d'impactes com en el coneixement i en la capacitació, no ens ha de sorprendre. A més, els impactes a curt termini són més fàcils de mesurar, mentre que els beneficis a llarg termini són més difícils i, de vegades, un repte obert. El cercle d'influència esmentat anteriorment no només afecta els investigadors, sinó també als avaluadors, ja que a mesura que es volen captar impactes més diversos i més allunyats de la recerca original, es fa més difícil d'atribuir-los directament a una recerca individual, sense que incorpori la col·laboració d'altra recerca o d'altres factors. No hi ha cap manual que descrigui com es trasllada la recerca cap a la seva aplicació, ja que aquesta forma part d'un món amb múltiples interconnexions i on el coneixement flueix fàcilment.

La participació dels investigadors amb usuaris potencials de la recerca va resultar un factor afavoridor de diferents impactes. Aquesta conclusió no és exclusiva d'aquesta tesi sinó que ve recolzada per altres estudis, per exemple en l'àmbit acadèmic,[75] de la indústria,[39] o de l'aplicació

en l'àmbit sanitari.[78] Involucrar la participació en diverses etapes del procés de recerca afavoreix una comunicació més efectiva, una recerca que incideix en les necessitats dels usuaris, i una major confiança. En general, promou decisions més legítimes, sensibles i rellevants, és a dir, una recerca més orientada a l'impacte. Aquests resultats no semblen ser excepcionals, ja que l'impacte final de la recerca es veu influenciat pel grau d'utilització dels coneixements obtinguts. Per tant, les relacions i xarxes personals dels IPs i de la resta de membres dels equips de recerca són un factor clau per a la translació d'aquesta.

D'altra banda, aspectes procedimentals com ara projectes orientats a la missió (orientats a aplicacions pràctiques) van associar majors beneficis socials. Això es pot interpretar com un pensament estratègic per part dels investigadors, en el sentit que consideren els 'mecanismes' potencials que poden millorar l'impacte de la seva recerca, és a dir, les vies que han d'acabar portant la recerca cap a una aplicació i un benefici social. En canvi, projectes destinats a omplir certes llacunes de coneixement, afavorien impactes acadèmics (coneixement i capacitació). És a dir, en el primer cas, la 'motivació' per fer la recerca podríem dir que és 'externa', ja que es busca els beneficis més enllà de l'àmbit de la recerca, mentre que en el segon cas, la consideraríem 'interna', quan es busca beneficis dins de la mateixa activitat.[79]

Amb aquests resultats s'ha volgut, d'una banda, retre comptes de la recerca que es fa a la UOC i, de l'altra, no només s'ha volgut quantificar els impactes sinó també, com podem donar pes a la nostra recerca per tal d'enfortir la capacitat humana i tècnica dels grups d'investigadors, d'aportar millores a la presa de decisions, i als beneficis econòmics i socials. Els factors trobats en aquesta tesi poden tenir-se en compte per part de decisors per tal d'ajudar a maximitzar l'impacte de la recerca, o a entendre el perquè del nombre de projectes que han aconseguit impactes. Afavorir un període de temps suficientment llarg, la participació d'usuaris potencials de la recerca o donar suport a projectes destinats a omplir certes llacunes de coneixement, pot ajudar a aconseguir impactes acadèmics. Així mateix, la participació d'usuaris potencials de la recerca o projectes orientats a la missió, pot afavorir impactes polítics o administratius, socials i econòmics. A més, els decisors han de tenir en compte que el fet d'aconseguir impactes acadèmics no és un factor determinant per aconseguir impactes més socials.

LIMITACIONS DELS ESTUDIS

Validesa d'un procés d'avaluació ex-ante de la recerca

En primer lloc, destacar que l'anàlisi s'ha fet sobre un procés molt específic d'avaluació de la recerca, i per tant cal anar amb cautela al interpretar els resultats d'aquesta tesi a altres sistemes. L'anàlisi es va realitzar en un procés de revisió per consemblants en què l'ocultació dels equips de recerca i les seves institucions només s'apliquen en la primera fase d'avaluació. En canvi, la majoria dels estudis que avaluen la validesa dels processos d'avaluació en relació a l'anonimització, estan realitzats amb l'avaluació d'articles per a revistes científiques i amb la comparació d'un procés completament anonimitzat amb un altre no anonimitzat.[29,80]

En segon lloc, cal tenir en compte la naturalesa no experimental de l'estudi. No s'ha utilitzat cap grup de comparació, ja que les dades provenen d'una anàlisi retrospectiva del procés i per tant l'experimentació comparativa podria millorar la validesa interna. Tampoc no s'ha pogut comprovar si els revisors consideraven les propostes completament anonimitzades.

En tercer lloc, atès que el qüestionari no incloia una pregunta específica i obligatòria que demanés quins eren els motius per fer un canvi, només s'ha pogut analitzar el camp obert que descriu comentaris addicionals en la segona fase, de manera que es van obtenir resultats indirectament. S'ha de considerar que en gairebé la meitat de les avaluacions que van canviar en la segona avaluació, el revisor no va incloure cap comentari i, com a tal, en aquests casos, va ser impossible saber quina era la justificació del canvi. A més, en un 5% dels casos, els comentaris que es van fer semblaven incongruents amb un canvi en la direcció oposada de l'avaluació. Tot i això, els resultats obtinguts en aquesta anàlisi tenen una coherència que no es pot obviar.

En quart lloc, en els estudis d'avaluació, sovint es recomana triangular i aprofundir els resultats a partir de mètodes i fonts múltiples per reforçar l'avaluació i reduir biaixos. Per aquest motiu, es van utilitzar mètodes quantitatius per analitzar el canvi i els seus factors, però es va necessitar una anàlisi qualitativa complementària per validar els resultats i respondre a preguntes sobre el perquè del canvi.

I, finalment, el nostre repte va ser analitzar diferents anualitats amb característiques molt diferents, i que podrien implicar característiques molt diverses, tot i que tots siguin de ciències de la salut. A més, al llarg de les diferents edicions s'han anat realitzant modificacions en el

procés (principalment en edicions antigues) que desconeixem si podrien afectar algun dels nostres resultats.

Validesa d'una eina d'avaluació ex-post de la recerca

El repte d'aquesta tesi ha estat analitzar una àmplia gamma de projectes que provenen de disciplines molt diferents: ciències socials, humanitats, ciències de la salut i TIC, i avaluar la validesa d'una eina que avaluï els seus diversos impactes en la societat. S'ha tractat de desenvolupar un instrument adaptable i variat utilitzant una gran diversitat d'indicadors per tal d'incloure el màxim possible d'impactes potencials, alhora que limitant la seva presència i aplicabilitat a alguna de les disciplines. Tanmateix, a causa de les diferències 'culturals' entre disciplines, no podem garantir que els IPs de diferents àrees de coneixement tinguin una comprensió homogènia de 'l'impacte de la recerca'. De fet, la diversitat en les respostes dels experts consultats per avaluar la rellevància dels indicadors del qüestionari suggereix el contrari. Per aquesta raó, una primera anàlisi de context, tal com es descriu en la literatura,[20] podria ajudar a decidir quins indicadors del qüestionari s'han d'incloure o eliminar en futurs estudis.

Cal destacar algunes limitacions pròpies dels estudis de validesa del contingut. D'una banda l'avaluació que fan els experts a vegades s'ha considerat subjectiva i sotmesa a l'experiència del mateix expert. Per aquest motiu s'ha volgut incloure experts de totes les disciplines analitzades, així com l'oportunitat que aquests poguessin suggerir (com ha estat el cas) qualsevol altre indicador que es considerés rellevant. El fet que els mateixos IPs, en el moment que contestaven el qüestionari tinguessin un camp obert on poder afegir altres impactes no inclosos entre els indicadors aportats, ajuda també a minimitzar aquesta limitació. El nombre d'experts requerits s'ha determinat en la majoria dels estudis com arbitrari, tot i que a mesura que augmenta el seu nombre, disminueix la probabilitat d'acord per l'atzar. En el desenvolupament d'eines d'avaluació, la validesa del contingut és un pas crític i un mecanisme per vincular conceptes abstractes a indicadors mesurables, però s'ha de considerar com un primer pas per estudiar la validesa de l'eina, ja que també seria recomanable incloure la validesa de constructe i la de criteri. L'evidència sobre propietats mètriques de les eines d'avaluació de l'impacte de la recerca són escasses,[40] limitant de la mateixa manera, la comparació d'aquests resultats amb la d'altres estudis. Per últim, són molt pocs els estudis que examinen les propietats mètriques de les eines d'avaluació de l'impacte de la recerca.[40] Per tant, la validesa externa de l'instrument, que és una mesura de rellevància, no s'ha pogut mesurar, tot i que hem de considerar que en el cas de l'avaluació de l'impacte de la recerca, aquesta pot ser rellevant en alguns contextos, encara que

no sigui generalitzable per altres. Per tant, el problema fonamental és la manca d'estàndards absoluts (anàlegs als estàndards físics per a mesuraments primaris, com ara el temps i la durada) per mesurar la qualitat de la investigació.

Respecte a les limitacions sobre l'estudi de fiabilitat, l'anàlisi d'aquesta es podria veure afectada per les característiques de la mostra. El fet que la majoria dels investigadors que van respondre el qüestionari pertanyin a ciències socials, limitaria que impactes relacionats amb la transferència, la comercialització i la innovació, siguin menys probables de produir-se, i per tant, l'anàlisi de la fiabilitat implicaria una menor consistència interna. També, el coeficient alfa i l'estudi de la consistència interna ha estat sovint criticat com a concepte i per raons estadístiques.[81] Però en aquesta tesi s'ha utilitzat de manera descriptiva, com un indicador de la coherència dels diferents ítems i pel que podria ser el fet d'eliminar ítems irrellevants.

Un altre punt són les limitacions relacionades amb els estudis d'impacte. La dificultat més important d'aquest tipus d'avaluació és intentar mesurar un constructe tan complex com el de l'impacte de la recerca. Les recomanacions internacionals sobre estudis d'avaluació d'impacte en la recerca aconsellen l'ús d'una combinació de mètodes i la seva triangulació per aconseguir resultats robustos i exhaustius.[20] No obstant això, com que l'enfocament principal de l'estudi ex post era la validesa de l'eina, no es va explorar la triangulació dels resultats amb altres mètodes. És a dir, els resultats del qüestionari presentats en aquesta tesi es poden considerar una primera fase exploratòria de l'impacte de la recerca en la UOC, els quals es podrien utilitzar en futurs estudis per seleccionar projectes que necessitin una anàlisi més profunda i detallada. També, el fet que siguin els mateixos investigadors els que responen el qüestionari implica un cert nivell de subjectivitat quan els IPs atribueixen un nivell d'impacte als seus projectes, cosa que podria fer que s'obtinguin resultats esbiaixats cap a la sobreestimació. En cap moment s'ha demanat als IPs que validin la informació que donen en el qüestionari aportant documents o dades que ho demostrin, fet que implicaria una major càrrega tant per als IPs com per als tècnics avaluadors, però que semblaria necessària per validar els resultats. Tot i això, Hanney et al.[14] demostren que els investigadors no exageren rutinàriament els impactes de la seva recerca, almenys en estudis similars a aquest, on les respostes donades no implicaven cap 'recompensa' o finançament futur. A més, cal dir que hi ha factors en l'estructura i l'organització dels equips de recerca que poden ser determinants de la productivitat i que no han estat incorporats en aquesta anàlisi, com són els diversos graus de dedicació a la recerca, a la docència o a l'assistència sanitària entre els membres dels equips de recerca, com també la promoció de joves investigadors, l'atracció d'investigadors postdoctorals o les bases interdisciplinàries dels

equips. Per mesurar el seu efecte hauria de dur-se a terme un estudi més complex (amb variables i nivells múltiples) i en profunditat dels equips de recerca i de les institucions investigadores.

Per últim, es necessita una mida de mostra correcta en els estudis de validesa. El requisit essencial és que la mostra sigui representativa de la població de la qual es dibuixa. En l'avaluació ex-post, tot i que la taxa de resposta va ser inferior a l'esperada, el 43% es troba dins del rang normal de qüestionaris en línia.[82] No obstant això, els investigadors d'arts i humanitats van estar poc representats. Una possible raó és que els investigadors no són plenament conscients dels impactes que es poden atribuir a la seva recerca. Una altra raó podria ser la creença que els estudis d'avaluació d'impacte de la recerca no poden proporcionar dades valuoses sobre com la recerca d'arts i humanitats genera impacte,[83] ja que part de la recerca no està relacionada amb una practicitat mesurable (publicacions, llibres o capítols) sinó amb canvis difícilment mesurables a escala social. Tot i aquests motius, s'han utilitzat amb èxit qüestionaris per mesurar l'impacte de la recerca en arts i humanitats,[18] i l'anàlisi dels casos del REF2014 va demostrar que la recerca en arts i humanitats proporcionava orientació i expertesa i que fàcilment podia ser utilitzada com evidència en debats públics, en la creació de polítiques i en l'aprenentatge institucional.[48]

IMPLICACIONS PER A LA RECERCA

Les complexitats dels estudis il·lustren que encara hi ha molt per aprendre en aquests camps sobre l'avaluació de la recerca.

Pel que fa a l'avaluació ex-ante, si bé validar el procés d'avaluació és completament necessari, també és important els resultats obtinguts en aquest procés. Una manera de validar aquest seria avaluant si els resultats són capaços de discernir les propostes de 'bona' qualitat. El significat de 'bona' qualitat no està estandarditzat, tot i que si fem cas del que diu la literatura científica es podria considerar com aquella recerca que és innovadora, interdisciplinària i aplicada.[25] Potser una manera seria preguntant als avaluadors no només la valoració de la proposta, sinó també, la confiança per ficar aquesta valoració. D'altra banda, una altra manera de validar el procés seria investigant si el procés és un bon pronòstic de l'èxit futur, i això hauria de ser reconegut, capturat i utilitzat per millorar la presa de decisions, ja que una de les crítiques en les revisions per consemblants és la manca de credibilitat en la seva fiabilitat predictiva. S'ha d'analitzar la relació entre les avaluacions ex-ante amb els impactes futurs d'aquesta recerca. Per últim, en

aquesta tesi només s'ha tingut en compte les primeres fases del procés d'avaluació. Quedaria pendent de validar l'última fase de l'avaluació ex-ante, aquella on les millors propostes són avaluades per un comitè ad-hoc. La investigació s'hauria de concentrar en la composició dels panells, incloent-hi l'experiència dels avaluadors, els seus camps d'expertesa, i com es prenen les decisions. Per últim, també s'hauria de considerar en futures investigacions que la qualitat dels resums de l'avaluació per als investigadors serveix perquè els investigadors percebin la validesa del procés d'avaluació.

Quant a l'estudi ex-post, resten per fer altres comprovacions, com podria ser la comparació de les opinions dels IPs amb les d'altres consemblants. És a dir, la valoració externa de l'impacte de la recerca en els diversos àmbits com una prova de la validesa de criteri. També seria recomanable, que a part dels experts utilitzats per a la validació dels ítems del qüestionari, es tinguessin en compte altres perspectives com la dels usuaris potencials de la recerca i així poder augmentar la robustesa social de la selecció dels indicadors i proporcionar un conjunt equilibrat de perspectives. Pel que fa a l'estudi d'impacte, seria útil aprofundir l'anàlisi amb la realització d'estudis de casos que permetin identificar el rang d'impactes aconseguits, les diverses vies de translació implicades per aconseguir aquests impactes i fins a quin punt es poden conceptualitzar i associar aquestes vies amb la naturalesa de les diferents recerques. Això ha de permetre potenciar la validesa de les troballes i reforçar els resultats trobats per tal de comprendre d'una manera més completa els diferents factors i interaccions que contribueixen a la translació de la recerca cap a l'impacte. Estudis a llarg termini i amb més aprofundiment podrien donar una visió més rica dels retorns socials de la recerca.

CONCLUSIONS

Les avaluacions realitzades han permès donar valor a dues eines i processos d'avaluació en concepte de validesa. L'anàlisi de la validesa de diferents enfocaments d'avaluació de la recerca (ex ante i ex post) pot contribuir de manera important a la credibilitat i acceptabilitat dels processos i de les eines d'avaluació de la recerca. Això ha ajudat a estimular i ampliar la reflexió sobre com desenvolupar una avaluació de la recerca metodològicament més robusta. D'una banda, els procediments de revisió per consemblants que faciliten inicialment l'enfocament només en els continguts del projecte de recerca (projectes anonimitzats) constitueixen una manera prometedora i vàlida de reduir alguns dels biaixos comuns descrits en la literatura sobre les característiques dels investigadors o de les disciplines i les decisions que es prenen per

finançar la recerca. D'altra banda, es proposa un qüestionari d'avaluació de l'impacte de la recerca, fàcil d'usar, i que permet mesurar una gran varietat de beneficis, oferint una bona coherència interna i una moderada validesa en el nostre context.

L'avaluació és una pràctica imperfecta. Tanmateix, fins i tot amb les seves limitacions, els enfocaments d'avaluació de la recerca poden proporcionar una informació molt valuosa per a la presa de decisions, sempre que es continuïn tenint en compte alguns reptes que poden limitar la seva validesa i, per tant, la seva credibilitat.

- La importància de la competència dels experts implicats en el procés d'avaluació és fonamental. Normalment, un expert es defineix com una persona que representa el contingut d'interès en el camp de la recerca o en la metodologia de l'avaluació. Una selecció precisa i la concordança dels avaluadors o revisors d'acord amb els seus coneixements pot ajudar a augmentar la validesa del procés.
- La validesa de l'eina d'avaluació de la recerca es pot millorar significativament quan es considera en context. És a dir, s'han de considerar els factors contextuals i analitzar el procés d'avaluació en funció de cada situació. Aquest és un punt clau del perquè no té sentit aplicar un model d'avaluació únic a escala global.
- Relacionat amb el punt anterior, també s'ha de tenir en compte que els procediments d'avaluació i els seus resultats han de ser acceptats pels grups d'interès. Disposar d'instruments i procediments sistemàtics i detallats augmenta la confiança d'aquells que els han d'utilitzar i millora les possibilitats de replicació.
- És molt important assegurar definicions clares dels conceptes explorats, puix que la manca de claredat pot amenaçar la validesa del procés. Els conceptes explorats s'han d'entendre de manera homogènia entre tots els avaluats.
- La definició de validesa interna afirma que les variacions en la variable dependent, són el resultat de variacions en les variables independents, i no pas d'altres factors confusors externs. Per aquest motiu, és important identificar la gamma completa de factors que es poden percebre com a beneficis o barreres per a un resultat determinat. Aquests beneficis i barreres identificats es poden incorporar a les eines per ajudar a comprendre el perquè del procés.
- Qualsevol avaluació es veu afectada per la qualitat i l'abast de les dades. La qualitat dels resultats és important i pot afectar la seva validesa interna. Afegir un apartat de comentaris per a una major justificació de les respostes, o preguntes addicionals per afegir informació sobre qualsevol aspecte no inclòs en els instruments d'avaluació,

ajuda a millorar la validesa dels instruments i garanteix que tots els resultats siguin més creïbles.

En conclusió, analitzar la validesa dels mètodes d'avaluació de la recerca és una mirada cap endavant que pot ajudar a entendre i millorar el procés i la seva credibilitat i, per tant, la utilitat de les mateixes avaluacions. Els exercicis d'avaluació han de tenir un objectiu específic, un coneixement del context i resoldre preguntes concretes i precises que permetin oferir respostes a la presa de decisions, així com implicar a les parts interessades en el moment de la discussió. La reflexió sobre el desenvolupament de mètodes vàlids i fiables pot dur a una millora dels processos de priorització, selecció i seguiment de la recerca segons els interessos dels actors socials que hi donen suport.

Referències

1. Ofir, Z., Schwandt, T., Duggan, C. & McLean, R. *Research Quality Plus (RQ+): A Holistic Approach to Evaluating Research*. (2016).

2. Guthrie, S., Wamae, W., Diepeveen, S., Wooding, S. & Grant, J. *Measuring Research: A Guide to Research Evaluation Frameworks and Tools*. (RAND Corporation, 2013).

3. Chalmers, I. *et al.* How to increase value and reduce waste when research priorities are set. *Lancet* **383,** 156–165 (2014).

4. ESF Member Organisation Forum on Evaluation of Publicly Funded Research. *Evaluation in Research and Research Funding Organisations: European Practices*. (2012).

5. OECD. *Manual de Frascati 2015: Guía para la recopilación y presentación de información sobre la investigación y el desarrollo experimental*. *FECYT* (2015). doi:10.1787/9789264310681-es

6. Foss Hansen, H. & Danmark. Forsknings- og Innovationsstyrelsen. *Research evaluation - methods, practice and experience*. (Danish Agency for Science, Technology and Innovation, 2009).

7. European Commission. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Strengthening the foundations of Smart Regulation – improving evaluation. 11 (2013).

8. Guthrie, S., Guerin, B., Wu, H., Ismail, S. & Wooding, S. Alternatives to Peer Review in Research Project Funding. (2013).

9. Shepherd, J., Frampton, G. K., Pickett, K. & Wyatt, J. C. Peer review of health research funding proposals: A systematic map and systematic review of innovations for effectiveness and efficiency. *PLoS One* **13,** 1–26 (2018).

10. Gurwitz, D., Milanesi, E. & Koenig, T. Grant application review: the case of transparency. *PLoS Biol.* **12,** e1002010 (2014).

11. Donovan, C., Butler, L., Butt, A. J., Jones, T. H. & Hanney, S. Evaluation of the impact of National Breast Cancer Foundation-funded research. *Med. J. Aust.* **200,** 214–218 (2014).

12. *Annual impact report 2014-2015*. (Alberta Innovates – Health Solutions, 2014).

13. Reed, R. L., Kalucy, E. C., Jackson-Bowers, E. & McIntyre, E. What research impacts do Australian primary health care researchers expect and achieve? *Heal. Res. Policy Syst.* **9,** 40 (2011).

14. Hanney, S., Watt, A., Jones, T. H. & Metcalf, L. Conducting retrospective impact analysis to inform a medical research charity's funding strategies: the case of Asthma UK. *Allergy Asthma. Clin. Immunol.* **9,** 17 (2013).

15. Milat, A. J. *et al.* Policy and practice impacts of applied research: a case study analysis of the New South Wales Health Promotion Demonstration Research Grants Scheme 2000-2006. *Heal. Res. Policy Syst.* **11,** (2013).

16. Bennett, J. W., Kelley, T. G. & Maredia, M. K. Integration of environmental impacts into ex-post assessments of international agricultural research: Conceptual issues, applications, and the way forward. *Res. Eval.* **21,** 216–228 (2012).

17. Molas-Gallart, J. Research evaluation and the assessment of public value. *Arts Humanit. High. Educ.* **14,** 111–126 (2015).

18. Levitt, R., Celia, C. & Diepeveen, S. Assessing the Impact of Arts and Humanities Research at the University of Cambridge. Technical Report. *RAND Corp.* 104 (2010).

19. Donovan, C. & Gulbrandsen, M. Introduction: Measuring the impact of arts and humanities research in Europe. *Res. Eval.* **27,** 285–286 (2018).

20. Adam, P. *et al.* ISRIA statement: Ten-point guidelines for an effective process of research impact assessment. *Heal. Res. Policy Syst.* **16,** (2018).

21. Gray, J. A. M. Evidence based policy making. *BMJ* **329,** 988–989 (2004).

22. Cronbach, L. J. (Lee J. *Essentials of psychological testing*. (Harper & Row, 1990).

23. Messick, S. Validity. in *Educational Measurement* (ed. Linn, R. L.) 13–103 (American Council on Education/ Macmillan Publishers, 1989).

24. Kane, M. T. An argument-based approach to validity. *Psychol. Bull.* **112,** 527–535 (2006).

25. Ismail, S., Farrands, A. & Wooding, S. Evaluating Grant Peer Review in the Health Sciences: A review of the literature. *Eval. Grant Peer Rev. Heal. Sci. A Rev. Lit.* (2009). doi:10.7249/tr742

26. Guthrie, S., Ghiga, I. & Wooding, S. What do we know about grant peer review in the health sciences? *F1000Research* **6,** 1335 (2017).

27. Mayo, N. E. *et al.* Peering at peer review revealed high degree of chance associated with funding of grant applications. *J. Clin. Epidemiol.* **59,** 842–848 (2006).

28. Garfield, E., Cole, S. & Rubin, L. Refereeing and Peer Review. Part 4. Research on the Peer Review of Grant Proposals and Suggestions for Improvement.

29. Baggs, J. G., Broome, M. E., Dougherty, M. C., Freda, M. C. & Kearney, M. H. Blinding in peer review: the preferences of reviewers for nursing journals. *J. Adv. Nurs.* **64,** 131–138 (2008).

30. Triggle, C. R. & Triggle, D. J. What is the future of peer review? Why is there fraud in science? Is plagiarism out of control? Why do scientists do bad things? Is it all a case of: "all that is necessary for the triumph of evil is that good men do nothing"? *Vasc. Health Risk Manag.* **3,** 39–53 (2007).

31. BUDDEN, A. *et al.* Double-blind review favours increased representation of female authors. *Trends Ecol. Evol.* **23,** 4–6 (2008).

32. Rossiter, M. W. The Matthew Matilda Effect in Science. *Soc. Stud. Sci.* **23,** 325–341 (1993).

33. Wessely, S. Peer review of grant applications: what do we know? *Lancet* **352,** 301–305 (1998).

34. Tamblyn, R., Girard, N., Qian, C. J. & Hanley, J. Assessment of potential bias in research grant peer review in Canada. *Cmaj* **190,** E489–E499 (2018).

35. Johnston, J. M. & Pennypacker, H. S. J. Strategies and tactics of behavioral research (3rd ed.). *Strategies and tactics of behavioral research (3rd ed.).* (2009).

36. Black, J. A. (James A. & Champion, D. J. *Methods and issues in social research*. (Wiley, 1976).

37. Lehner, P. N. *Handbook of ethological methods*. (Cambridge University Press, 1998).

38. Nunnally, J. C. & Bernstein, I. H. *Psychometric theory*. (McGraw-Hill, 1994).

39. Bland J; Altman D. Statistics notes: Cronbach's alpha. *BMJ* **314,** 275 (1997).

40. Aymerich, M. *et al.* Measuring the payback of research activities: A feasible ex-post evaluation methodology in epidemiology and public health. *Soc. Sci. Med.* **75,** 505–510 (2012).

41. Health Economics Research Group, B. U., Office of Health Economics & RAND Europe. Medical Research: What's it worth? Estimating the economic benefits of research in the UK. 1–108 (2008).

42. Kunac, D. L., Reith, D. M., Kennedy, J., Austin, N. C. & Williams, S. M. Inter- and intra-rater reliability for classification of medication related events in paediatric inpatients. *Qual. Saf. Heal. Care* **15,** 196–201 (2006).

43. Milat, A. J., Bauman, A. E. & Redman, S. A narrative review of research impact assessment models and methods. *Health Res. Policy Syst.* **13,** 18 (2015).

44. Universitat Oberta de Catalunya. *Strategic Plan Stage II 2017-2020*. (2017).

45. San Francisco Declaration on Research Assessment (DORA). Available at: https://sfdora.org/. (Accessed: 17th December 2018)

46. IAU HESD Cluster | HESD - Higher Education for Sustainable Development portal. Available at: http://iau-hesd.net/en/contenu/4648-iau-hesd-cluster.html. (Accessed: 17th December 2018)

47. Wooding, S., Nason, E., Starkey, T., Hanney, S. & Grant, J. *Mapping the impact: Exploring the payback of arthritis research*. (RAND Corporation, 2010).

48. Higher Education Funding Council of England *et al. The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*. (2015).

49. Peterson, R. A. A Meta-Analysis of Cronbach's Coefficient Alpha. *J. Consum. Res.* **21,** 381 (1994).

50. ESF. *ESF survey analysis report on peer review practices*. (2011).

51. Ioannidis, J. P. A. More time for research: Fund people not projects. *Nature* **477,** 529–531 (2011).

52. Gluckman, P. Which science to fund: time to review peer review? 11 (2012).

53. Abdoul, H. *et al.* Peer review of grant applications: criteria used and qualitative study of reviewer practices. *PLoS One* **7,** e46054 (2012).

54. Bhattacharjee, Y. NSF's 'Big Pitch' Tests Anonymized Grant Reviews. *Science (80-.).* **336,** 969–970 (2012).

55. Merton, R. K. The Matthew Effect in Science: The reward and communication systems of science are considered. *doi.org* **159,** 56–63 (1968).

56. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: A randomized experiment. *Science (80-.).* **341,** 647–651 (2013).

57. Marsh, H. W., Jayasinghe, U. W. & Bond, N. W. Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *Am. Psychol.* **63,** 160–168 (2008).

58. van der Lee, R. & Ellemers, N. Gender contributes to personal research funding success in The Netherlands. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 12349–53 (2015).

59. Tricco, A. C. *et al.* Strategies to Prevent or Reduce Gender Bias in Peer Review of Research Grants: A Rapid Scoping Review. *PLoS One* **12,** e0169718 (2017).

60. Bornmann, L., Mutz, R. & Daniel, H.-D. Gender differences in grant peer review: A meta-analysis. *J. Informetr.* **1,** 226–238 (2007).

61. Kaatz, A., Gutierrez, B. & Carnes, M. Threats to objectivity in peer review: the case of gender. *Trends Pharmacol. Sci.* **35,** 371–3 (2014).

62. Volker, B. & Steenbeek, W. No evidence that gender contributes to personal research funding success in The Netherlands: A reaction to van der Lee and Ellemers. *Proceedings of the National Academy of Sciences of the United States of America* **112,** E7036–E7037 (2015).

63. Marsh, H. W., Jayasinghe, U. W. & Bond, N. W. Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model. *J. Informetr.* **5,** 167–180 (2011).

64. Turner, S., Davidson, P., Stanton, L. & Cawdeary, V. Features of successful bids for funding of applied health research: a cohort study. *Heal. Res. policy Syst.* **12,** 54 (2014).

65. Clark, J. & Horton, R. What is The Lancet doing about gender and diversity? *Lancet (London, England)* **393,** 508–510 (2019).

66. Reinhart, M. Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics* **81,** 789–809 (2009).

67. Jang, D., Doh, S., Kang, G.-M. & Han, D.-S. Impact of Alumni Connections on Peer Review Ratings and Selection Success Rate in National Research. *Sci. Technol. Hum. Values* **42,** 116–143 (2017).

68. Mastaglia, B., Toye, C. & Kristjanson, L. J. Ensuring content validity in instrument development: challenges and innovative approaches. *Contemp. Nurse* **14,** 281–91 (2003).

69. Kalucy, E. C., Jackson-Bowers, E., McIntyre, E. & Reed, R. The feasibility of determining the impact of primary health care research projects using the Payback Framework. *Heal. Res. Policy Syst.* **7,** 11 (2009).

70. Guthrie, S. *et al. A 'DECISIVE' approach to research funding: Lessons from three Retrosight studies*. (RAND Corporation, 2016). doi:10.7249/RR1132

71. Wooding, S., Nason, E., Thompson-Starkey, T. G., Hanney, S. & Grant, J. Mapping the impact. (2009).

72. Wilsdon, J. *et al.* The Metric Tide Report of the Independent Review of the Role of Metrics in Research Assessment and Management. (2015). doi:10.13140/RG.2.1.4929.1363

73. Wooding, S. *et al.* Understanding factors associated with the translation of cardiovascular research: a multinational case study approach. *Implement. Sci.* **9,** 47 (2014).

74. *OECD Reviews of Regional Innovation: Catalonia, Spain 2010*. (OECD Publishing, 2010). doi:10.1787/9789264082052-en

75. RAND Europe. Mental Health Retrosight. (2013).

76. Wong, P. K. & Singh, A. Do co-publications with industry lead to higher levels of university technology commercialization activity? *Scientometrics* **97,** 245–265 (2013).

77. Solans-Domènech, M. *et al.* Impact of clinical and health services research projects on decision-making: A qualitative study. *Heal. Res. Policy Syst.* **11,** (2013).

78. Solans-Domènech, M. *et al.* Impact of clinical and health services research projects on decision-making: a qualitative study. *Health Res. Policy Syst.* **11,** 15 (2013).

79. Ryan, R. M. & Deci, E. L. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemp. Educ. Psychol.* **25,** 54–67 (2000).

80. Mulligan, A., Hall, L. & Raphael, E. Peer review in a changing world: An international study measuring the attitudes of researchers. *J. Am. Soc. Inf. Sci. Technol.* **64,** 132–161 (2013).

81. Sijtsma, K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* **74,** 107–120 (2009).

82. Shih, T. H. & Xitao, F. Comparing response rates from web and mail surveys: A meta-analysis. *Field methods* **20,** 249–271 (2008).

83. Molas-Gallart, J., D'Este, P., Llopis, O. & Rafols, I. Towards an alternative framework for the evaluation of translational research initiatives. *Res. Eval.* **25,** 235–243 (2016).

Annexos

ANNEX 1. PUBLICACIÓ #1 (AVALUACIÓ EX-ANTE)

Research Evaluation, 26(3), 2017, 181–189
doi: 10.1093/reseval/rvx021
Advance Access Publication Date: 30 May 2017
Article

OXFORD

Blinding applicants in a first-stage peer-review process of biomedical research grants: An observational study

Maite Solans-Domènech[1,2,3,*], Imma Guillamón[1,2], Aida Ribera[2,4], Ignacio Ferreira-González[4], Carme Carrion[3,5], Gaietà Permanyer-Miralda[1,4] and Joan M. V. Pons[1,2]

[1]Agency for Health Quality and Assessment of Catalonia (AQuAS), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [2]Epidemiology and Public Health Network (CIBER ESP), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [3]Faculty of Health Sciences, Universitat Oberta de Catalunya, Barcelona, Catalonia 08018, Spain, [4]Epidemiology Unit, Cardiology Service, Vall d'Hebron Hospital, Passeig de la Vall d'Hebron, 119-129, Barcelona, Catalonia 08035, Spain and [5]Laboratory of Translational Medicine and Decision Science (TransLab Research Group), Department of Medical Sciences, Faculty of Medicine, University of Girona, Girona, Catalonia 17071, Spain

*Corresponding author: E-mail: mtsolans@gencat.cat.

Abstract

To blind or not researcher's identity has often been a topic of debate in the context of peer-review process for scientific publication and research grant application. This article reports on how knowing the name and experience of researchers/institutions influences the qualification of a proposal. We present our experience of managing the peer-review process of different biomedical research grants. The peer-review process included three evaluation stages: first, blinded assessment; second, unblinded assessment by the same reviewer; and final, assessment of the better qualified proposals by an *ad hoc* committee. The change between the first (applicants blinded) and the second assessments (unblinded) for each evaluation and reviewer was evaluated. Factors associated with change were analysed, taking into account the characteristics of proposals, reviewers, and researchers. A qualitative content analysis of the reviewers' comments was also carried out to assess the reasons for change. The analysis of 5,002 evaluations indicated that in 18.5% of the evaluations (from 10.5 to 27.7% depending on the year of the edition), the reviewer changed the second assessment: either for better (11.9%) or worse (6.6%). Our findings also suggest that a change in the second assessment was highly correlated with a positive evaluation of the experience of the principal investigator or research team. With a change of 1 in 10 to 1 in 4 depending on the year of the edition, we believe that concealing the identity of researchers/institutions could help to focus exclusively on the proposal and reduce some of the common biases of the peer-review process in grant decisions.

Key words: grant peer review; financing; organized; double-blind method; qualitative research; quality control.

1. Introduction

Peer review is the most widespread system used to allocate research funds and to appraise scientific manuscripts for publication. Although imperfect, it is an accepted instrument for self-policing and ensuring quality in scientific research, having gained as much legitimacy in the scientific world as among the lay public (Gurwitz et al. 2014). Since the 80s, there has been a quest to improve the peer-review process in its robustness, fairness, and transparency.

As an object of scientific study, the peer-review process is only a few decades old. Most of the literature has focused on the peer-

review process regarding submissions to journals (Mayo et al. 2006). Blinding reviewers and/or authors of manuscripts has often been a topic of discussion (Wessely 1998; Regehr and Bordage 2006; Baggs et al. 2008; Mulligan et al. 2013). In principle, blinding can reduce bias against some researchers' characteristics (i.e. female or junior researchers) (Triggle and Triggle 2007; Budden et al. 2008) or proposal characteristics (i.e. untried, very innovative or disruptive, and interdisciplinary proposals) (Wessely 1998). Avoidance of any interpersonal conflict of interest is another argument supporting the blinding of identities (Baggs et al. 2008). Most of the evidence available related to the effect of blinding the author on the quality of reviews of manuscripts submitted to the editorial peer-review system demonstrates little or no effect (Justice et al. 1998; Van Rooye et al. 1998, 1999; Alam et al. 2011).

There has been less written on peer review of grant applications. One of the best-known studies of peer review of grant applications was that performed in the 70s to 80s at the request of the Committee on Science and Public Policy of the US National Academy of Sciences on the review process of the National Science Foundation (NSF). Experimentally, half of the reviewers received proposals that had been edited in an attempt to conceal the applicant's identity; the other half received copies identical to those but without concealing the applicants, as they had been submitted to the NSF. Results of the peer review in the NSF showed that an applicant's age and track record had little effect on the chances of getting a grant and that reviewers treated proposals from researchers at prestigious institutions no differently than proposals from workers at less prestigious institutions (Garfield 1987). Reviewers of blinded proposals were also asked whether the removal of title pages, list of references, budgets, or any other identifying information made the proposal more difficult to evaluate. The study found that it was difficult to conceal authorship because it made the proposal almost unreadable. Also, complete blinding often seems difficult to achieve (because of the many internal clues pointing towards authorship included in the articles) (Van et al. 1999). The Committee concluded that the blinding process of grant applications severely compromised the integrity of the proposals. Contrary to manuscript peer review, which is an ex-post research assessment, review of grant for funding (ex ante research assessment) was considered highly dependent on the principal investigator's (PI)/research team's ability to adequately implement the proposal. Another study that analysed the gap in success rates between different races and ethnicities showed a bias against black or Asian researchers, even after controlling for education, country of origin, training, employer characteristics, previous research awards, and publication record (Tabak and Collins 2011).

Since then, more literature has been added and many different biases have been reported in the peer-review process: age, institution, 'cronyism', discipline, gender, etc. A non-systematic review of the existing studies on peer review for awarding grants was published at the end of last century by Wessely (1998) which examined issues of equity, efficiency, and failure to promote the best science. A lack of reliability in the rankings of reviewers was, among other things, one of the main weaknesses considered. More recently the Cochrane Collaboration performed another more comprehensive and systematic review, examining the effects of peer review in awarding grants, taking into account different ways of screening, assigning, or masking submissions; different ways of eliciting internal or external opinions; different ways of carrying out procedures (single person or group); and different types of feedback given and revisions done of applications (Demicheli and Di Pietrantonj 2007). The authors of the review concluded that 'there is little empirical evidence of the effects of peer review in awarding grants' (importance, relevance, usefulness, soundness of methods, soundness of ethics, completeness, and accuracy) because they were unable to find comparative studies assessing the actual effect of peer-review procedures on the quality of the research funded. There was an urgent need for research to fill this gap and, as Wessely mentioned, the absence of controlled trials in this area of scientific decision making was ironic (Wessely 1998).

A recent call for greater transparency in reviewing grant applications once again mentioned both the low agreement between reviewers in their qualifications and also the more recent case of the NSF where reviewers faced with blinded proposals selected a different set of projects for funding than those chosen by reviewers with unblinded versions of the same proposal (Bhattacharjee 2012; Gurwitz et al. 2014). In any case, a recent survey has shown that blinding applicants is extremely rare among public research financing agencies, with only 4% of the organizations doing so (ESF 2011).

The peer-review process described here, and managed for more than 10 years now, consists in a first blinded assessment followed by a second unblinded assessment. So, the goal of this study is to analyse whether concealing the identity of researchers and their institutions from peer reviewers in the first assessment stage of a research project changes the reviewer's assessment when the name of the PI/research team, their experience, and the institution they represent is revealed in a second stage. Specifically, our intention was to quantify the change, its direction, and to know whether there are any factors associated and the reasons that led to it.

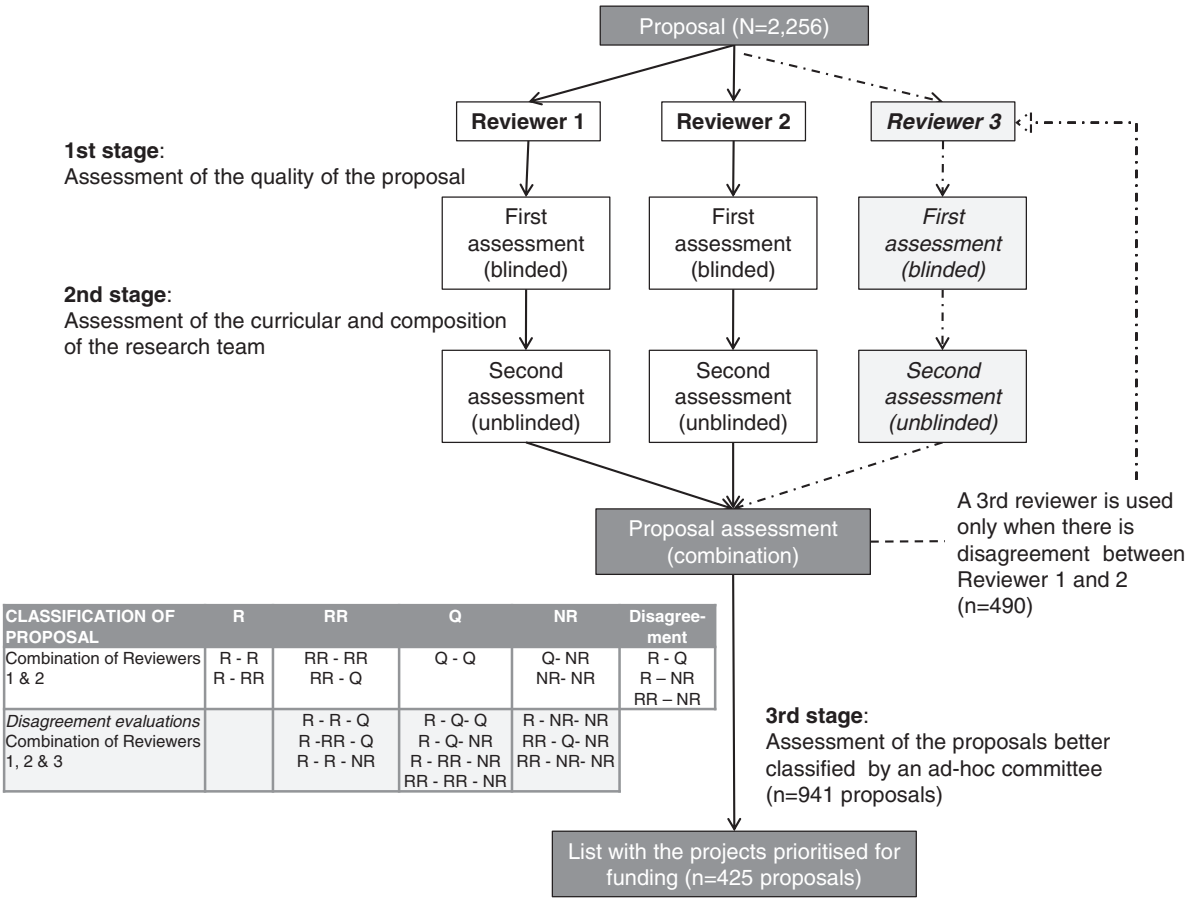
2. Methodology

A retrospective observational study of the peer-review process carried out to evaluate different biomedical research grant applications. The sample consisted of all research proposals (N = 2,256) presented in 14 annual calls from 2002 to 2015 (2001–2014 editions). Projects were evaluated by 1,475 international reviewers (about two to three reviews for each proposal, on average 2.2). Overall, 5,002 evaluations were conducted. First, we must describe the assessment process.

2.1 Peer-review process overview

Since 2001, the Agency for Health Quality and Assessment of Catalonia (AQuAS) has been in charge of managing the peer-review process of different annual research calls for biomedical research projects. Unlike other research calls, the main topic of interest in these calls changes each year. From the outset, proposals have had to be submitted in English, and the peer-review process, done in three stages, only has non-Spaniard reviewers who independently assess the scientific quality of projects (Fig. 1). Reviewers are mainly selected by searching through medical literature, or from scientific societies, editorial journals, or reviewer repositories.

The first assessment stage (blinded) begins when the details of all researchers linked to a research proposal and the institutions they represent are concealed. Researchers' names, manuscripts' titles, volume, and pages (but not journal identification or year of publication) are also suppressed from self-references in the text and in the bibliographic section. Therefore, the assessment only focuses on the content of the research proposal and performed by using a



CLASSIFICATION OF PROPOSAL	R	RR	Q	NR	Disagree-ment
Combination of Reviewers 1 & 2	R - R R - RR	RR - RR RR - Q	Q - Q	Q- NR NR- NR	R - Q R – NR RR – NR
Disagreement evaluations Combination of Reviewers 1, 2 & 3		R - R - Q R -RR - Q R - R - NR	R - Q- Q R - Q- NR R - RR - NR RR - RR - NR	R - NR- NR RR - Q- NR RR - NR- NR	

Figure 1. Workflow of the review process.
Note: R = recommended for funding; RR = recommended with reservations; Q = Questionable; NR = not recommended.

structured questionnaire with response categories on a Likert scale. At the end of the questionnaire, the reviewers are asked to state whether, in qualitative terms, the project should be recommended (R), recommended with reservations (RR), questionable (Q), or not recommended (NR) for funding. In the second stage (unblinded), reviewers once again use a structured questionnaire to evaluate the experience and track record of the PI, the research team, and the suitability of the institution where the research is planned. At this point, the reviewers have to give their second and definitive qualitative assessment of the project in the same terms (R, RR, Q, or NR). When there is a two-level disagreement in this second stage between two reviewers, the proposal is sent to a third reviewer for another independent evaluation performed in the same way. The time elapsed between the first and second assessments is approximately 1–2 weeks. The second stage begins when the AQuAS receives the first completed questionnaire. The two or more categories obtained from the assessments are combined for each proposal into one that allows the proposal to be classified in self-excluding categories (R, RR, Q, and NR) (see Fig. 1).

The third and final stage of the process is performed by an *ad hoc* committee composed of some of the international reviewers who meet over a 2-day period. These reviewers are in charge of assessing the best qualified proposals in the second stage (usually those classified as R, and sometimes also the RR's). The committee drafts a list of the projects prioritized for funding, taking into account the amount of funds available. This list is presented to the respective Scientific Advisory Committee, and the final decision is made by the board of trustees. The process ends when the proposals awarded grants are made public, and a report describing the entire peer-review process and the reviewers participating in it is published.

2.2 Quantitative analysis

The primary variable of analysis was the change made between the first (researchers/institutions blinded) and the second assessments (unblinded) for each evaluation and for each reviewer. The variable was categorized on an ordinal scale of 0 = no change, 1 = improvement (change for a better category), and 2 = worsening (change for a worse one). The relationship between the primary variable and the adequation of the PI/research team/institution was analysed with a chi-square test. The PIs/research teams/institutions were rated in the second-stage questionnaire by the reviewers according to their appropriateness to carry out the proposal and placed in one of four categories (strongly agree, agree, disagree, or strongly disagree). This variable was dichotomized for this study. Agreement between the first (blinded) and the second assessments (unblinded) was calculated using the weighted kappa statistic (k).

An adjusted multinomial logistic regression model was used to identify those factors associated with change with 'no change' as the reference category. Relative risk ratios (RRRs) were calculated

for improvement and for worsening. Several covariates were included in the data analyses as predictors, both at the level of proposals and PIs, and reviewers. At the level of proposals, 'year of the edition', 'research area of the proposal' identified by the applicant, 'requested grant sum', 'PI's gender', 'PI's age', and 'adequate experience of the research team' were taken into account. The maximum sum of the grant requested allowed for variations depending on the type of proposal; individual with a single group/institution coordinated between two groups or coordinated among three or more participating groups from different institutions. At the level of reviewers, the 'reviewer's gender', 'world region of the reviewer', and 'reviewer h-index' were included. The gender (and age) of the applicant was obtained from the forms completed on submission of the research project or in the reviewer acceptance form. If gender was not specifically identified, we conducted a manual search (using websites or corresponding addresses). In particular, Asian names were difficult to assign to a specific gender, so in those cases the gender variable was left blank. The h-index was determined for each reviewer using *Web of Science* and taking into account the year of the evaluation. The value of h is equal to the number of papers of the reviewer (N) that have N or more citations. The statistical significance was set at $P \leq 0.05$. Statistics were calculated using SPSS18.

2.3 Qualitative analysis

To analyse the reasons that indirectly influenced a change in the second assessment (unblinded), a qualitative content analysis was carried out. As the evaluation questionnaire does not have a specific field for documenting the reasons for a change, we analysed the open field for additional comments included in the second-stage evaluation form. The only cases examined were those in which a change was made and a comment written by the reviewer. Of the 922 evaluations that modified the assessment, half of them (n = 461) filled in the open field for additional comments. The content analysis was conducted in an inductive two-dimensional way: reasons for change and nature of the reason. First, one author read all the comments to get an overall understanding of the reviewers' comments and to extract the initial categories and subcategories. Then, the comments were classified into these categories and subcategories by all the authors. Secondly, we determined whether the justifications were of a positive, negative, or neutral nature. A justification was defined as positive when the reviewer described excellent aspects of the project; a negative justification meant that the characteristics evaluated did not seem appropriate or were absent; and a neutral character was considered when it was impossible to classify the justification as either positive or negative. The results were triangulated by two coders, who reached a consensus and discussed those cases classified as 'doubtful' within the multidisciplinary team.

3. Results

Table 1 displays a description of the different editions analysed. During the period 2001–2014, more than 101 million € have been distributed among 376 projects, with a success rate of 18.6% for the 2,256 proposals presented. Overall 1,475 international reviewers participated in this assessment process (some participated in more than one edition). A third reviewer to solve discordances was needed in 490 cases (9.8%).

3.1 Quantitative analysis

We analysed 5,002 evaluations. In most of them (81.5%) and after the PI/research teams/institution were unblinded, reviewers did not change their second assessment of the proposal, while in 18.5% (n = 922) cases, there was a change in the assessment: it was for better in 11.9% (n = 594) cases and for worse in 6.6% (n = 328). Depending on the year of the edition year, the percentage of change ranged from 10.5 to 27.7% (Table 1).

The association between an adequate experience of the PI/research team and the change in the second assessment was statistically significant ($P < 0.05$). The change was for the better when there was substantial agreement on the 'adequate experience' of the PI/research team, while it was for the worse when there was considerable disagreement.

Both the first (blinded) and the second (unblinded) assessments included similar rates of the four possible categories (R, RR, Q, and NR). A weighted Kappa statistic indicates a very good agreement ($k = 0.75$) between the first and second assessments.

In the adjusted multinomial model, a positive evaluation of the experience of the PI/research team/institution showed the strongest association with a positive change (RRR = 2.63; $P < 0.005$), and it was less likely to have a negative change (RRR = 0.25; $P < 0.005$). Earlier editions, from 2001 to 2007 and 2009 edition) were also found to be a factor associated with positive change, and there were only negative changes in the years 2002 and 2004. Multivariate analyses also showed that, compared to no change, a positive change was less likely to be present with reviewers coming from North America (RRR = 0.62; $P = 0.001$), in comparison to European reviewers. In contrast, when a budget of more than €300,000 was requested, there was less likelihood (RRR = 0.49; $P = 0.01$) of a negative change. Female PI was also a statistical significant factor associated with a negative change (RRR = 1.42; $P = 0.001$) with no change as reference category (Table 2). Other factors such as research area of the proposal, PI's age, reviewer's gender, or reviewer h-index had no association with any positive or negative change.

3.2 Qualitative analysis

There was an added comment in 50% (461) of the evaluations which changed the qualification of the project in the second assessment stage. Of these comments, 5% (n = 23) were inconsistent with the change, i.e. the reviewer improved the second assessment, although comments were unfavourable (n = 18), or he/she worsened the second assessment without a clear rationale for change (n = 5). That means that the justification could be 'positive', when it presented the strengths of the research team or the proposal, or 'negative', when the assessment implied its weaknesses.

The justification ranged in length from 2 ('Excellent team' [EvaluationID 90]) to 452 words. The themes emerging from the reviewers' comments fell into three main groups related to: (1) an evaluation, (2) a suggestion, or (3) a comment about the lack of information either related to the PI/research team or to the proposal. The comments that describe an evaluation (first group) referred to, first, the skills and experience of the research team and/or the PI that demonstrate whether they have an adequate background for carrying out the proposal, mostly measured by the publications of group members. Also, it was measured regarding multi-institutional collaboration, which was not only related to the need for more participating centres (i.e. to collect enough samples) but also to the aspect of how the multiple centres should be coordinated. Finally, it also

Table 1. Descriptive characteristics of the different editions

Year	2001	2002	2003	2004	2005	2006	2007	20008	2009	2010	2011	2012	2013	2014	Total
Number of proposals presented	107	80	105	260	200	81	159	151	244	100	205	317	147	100	**2,256**
Research area of proposal: number (%)															
Basic	49 (46)	38 (48)	32 (30)	132 (51)	139 (70)	26 (32)	66 (42)	47 (31)	128 (52)	40 (40)	119 (58)	162 (51)	81 (55)	27 (27)	**1,086 (48)**
Clinical	33 (31)	16 (20)	47 (45)	69 (27)	42 (21)	39 (48)	49 (31)	60 (40)	53 (22)	35 (35)	48 (23)	62 (20)	29 (20)	28 (28)	**610 (27)**
Epidemiology/health services	15 (14)	2 (3)	10 (10)	22 (8)	3 (2)	8 (10)	14 (9)	20 (13)	5 (2)	4 (4)	6 (3)	22 (7)	3 (2)	6 (6)	**140 (6)**
Combination	10 (9)	24 (30)	16 (15)	37 (14)	16 (8)	8 (10)	30 (19)	24 (16)	58 (24)	21 (21)	32 (16)	71 (22)	34 (23)	39 (39)	**420 (19)**
Type of proposal: number (%)															
Single institution	88 (82)	55 (69)	85 (81)	214 (82)	144 (72)	71 (88)	106 (67)	99 (66)	140 (57)	62 (62)	121 (59)	185 (58)	71 (48)	63 (63)	**1,504 (67)**
Coordinated	19 (18)	25 (31)	20 (19)	46 (48)	56 (28)	10 (12)	53 (33)	52 (34)	104 (43)	38 (38)	84 (41)	132 (42)	76 (52)	37 (37)	**752 (33)**
Number of international reviewers	52	34	52	169	201	92	145	131	134	111	105	257	96	59	**1,638**[a]
Number of proposals funded	21	20	26	30	35	28	25	26	20	30	29	42	44	43	**376**
Amount of money collected (million €)	3.9	3.8	3.3	7.7	6.9	7.1	7.0	6.1	6.3	8.7	8.0	11.3	11.2	10.3	**101.6**
Average funding per proposal (thousand €)	186.1	188.8	128.6	256.4	196.8	254.1	281.0	236.4	317.5	291.1	277.1	269.9	255.0	239.6	**243.2**
Proposals success rate[b] (%)	19.6	25.0	24.8	11.5	17.5	34.6	15.7	17.2	8.2	30.0	14.1	13.2	29.9	43	**18.6**
Percentage of evaluations that change the qualitative category (%)	26.2	24.4	24.7	23.7	27.7	17.3	21.9	15.3	20.9	11.3	10.5	12.7	13.5	11.1	**18.5**

Note: [a]Some of the reviewers participated in more than one edition.
[b]Proposal success rate was calculated dividing the number of proposals funded by the number of proposals presented.

included the composition and multidisciplinary nature of the team, which was described in relation to the different backgrounds, expertise, and specialization of its members. On the other hand, the assessment of a proposal was related to its strengths and weaknesses, such as relevance, quality, or importance, characteristics that were in fact part of the first-stage assessment. The second group (a suggestion) includes suggestions regarding the research team, which were mainly related to the necessity of additional expertise and suggestions regarding the proposal, including a wide array of aspects such as budget, hypothesis, planning. Finally, the third group, regarding a lack of information, was considered when reviewer justification described that the proposal did not possess enough information to evaluate one specific aspect (either of the research team or the proposal). An overview of the results (and quotes) is shown in Table 3.

4. DISCUSSION

The results of this analysis indicate that in 18.5% of the evaluations, from 10 to 28% depending on the edition, the assessment of a research proposal was changed after the identity and experience of the researchers and their institutions were revealed. Therefore, most of the reviewers maintain their initial assessment of the blinded proposal. As expected, our findings also imply that the change in the second assessment was highly correlated with the evaluation of the appropriateness of the PI/research team's experience. This is not surprising, as the track record of the researchers and their institutions was the only new information received in the second stage. These results are also supported by the qualitative analysis.

The fact that almost 19% of the assessments were changed after the details of the researchers were made known opens a further discussion about the time and cost of blinding applicants. Is this percentage enough to continue insisting on and ensuring a researcher's anonymity for the first (blinded) assessment stage of proposals? It is not very common among agencies financing research to request that proposals be blinded as a recent survey has shown (ESF 2011), and it is even less common to ensure its effectiveness. However, there was no question in the survey regarding the blinding of applicants in the first-stage review process.

In the qualitative analysis of a reviewer's comments, the reasons for changing an assessment were mainly characterized by a positive or a negative evaluation of the PI/research team. The skills and experience of the research team and/or the PI, multi-institutional collaboration, or the composition and multidisciplinary nature of the team were factors that the reviewers commented on in the second phase of evaluation as a reason to change their assessment. Therefore, this played an important part in judging the feasibility of a proposal and supports the studies of the NSF which considered that blinding applicants compromised the integrity of a proposal. However, in almost half of the evaluations that changed in the second assessment, no comment was included by the reviewer and as such, in these cases, it was impossible to know what the justification was for the change. Moreover, in 5% of the cases, the comments made seemed to be incongruous with a change in the opposing direction of the assessment. Our results also seem to reaffirm the importance of evaluating a proposals' characteristics (quality, originality, methodology, innovation, etc.), as other studies have shown (Abdoul et al. 2012), because the assessments remained unchanged in the majority of cases.

Table 2. Factors predicting change between the first assessment (applicants blinded) and the second assessment (unblinded)

Factors	Number of evaluations (% total)	Positive change rate (% category)	Negative change rate (% category)	Adjusted model RRR (95% CI)	
				Positive change	Negative change
Edition year					
2001	214 (4.28)	41 (19.16)	15 (7.01)	**3.83 (1.91–7.68)**	1.45 (0.58–3.58)
2002	160 (3.20)	21 (13.13)	18 (11.25)	**2.55 (1.20–5.40)**	**2.47 (1.04–5.86)**
2003	231 (4.62)	33 (14.29)	24 (10.39)	**3.03 (1.53–6.03)**	2.16 (0.96–4.88)
2004	586 (11.73)	78 (13.31)	61 (10.41)	**2.35 (1.28–4.32)**	**2.42 (1.18–4.99)**
2005	451 (9.03)	96 (21.29)	29 (6.43)	**4.02 (2.19–7.37)**	1.20 (0.55–2.61)
2006	185 (3.70)	22 (11.89)	10 (5.41)	**2.33 (1.13–4.80)**	0.88 (0.34–2.25)
2007	356 (7.13)	56 (15.73)	22 (6.18)	**2.78 (1.46–5.32)**	1.48 8 (0.64–3.41)
2008	339 (6.79)	29 (8.55)	23 (6.78)	1.03 (0.47–2.24)	1.07 (0.42–2.68)
2009	540 (10.81)	74 (13.70)	39 (7.22)	**2.36 (1.29–4.34)**	1.94 (0.93–4.07)
2010	222 (4.44)	14 (6.31)	11 (4.95)	0.98 (0.45–2.17)	0.73 (0.28–1.88)
2011	457 (9.15)	29 (6.35)	19 (4.16)	0.94 (0.48–1.85)	0.80 (0.35–1.79)
2012	703 (14.07)	61 (8.68)	28 (3.98)	1.44 (0.78–2.67)	0.79 (0.37–1.70)
2013	326 (6.53)	25 (7.67)	19 (5.83)	1.20 (0.60–2.39)	1.19 (0.53–2.72)
2014	225 (4.50)	15 (6.67)	10 (4.44)	Ref.	Ref.
Research area of the proposal					
B	2,387 (47.79)	309 (12.95)	159 (6.66)	0.99 (0.62–1.57)	1.09 (0.61–1.98)
C	1,354 (27.11)	143 (10.56)	95 (7.02)	1.03 (0.80–1.32)	1.21 (0.84–1.73)
E	306 (6.13)	36 (11.76)	25 (8.17)	0.88 (0.65–1.18)	1.01 (0.68–1.51)
Combinations	948 (18.98)	106 (11.18)	49 (5.17)	Ref.	Ref.
Requested grant sum					
<€100,000	471 (9.44)	45 (9.55)	46 (9.77)	Ref.	Ref.
€100,000–€199,999	2,503 (50.15)	294 (11.75)	189 (7.55)	1.22 (0.84–1.76)	0.98 (0.66–1.45)
€200,000–€299,999	929 (18.61)	104 (11.19)	53 (5.71)	1.28 (0.84–1.95)	0.73 (0.45–1.19)
≥€300,000	1,088 (21.80)	150 (13.79)	40 (3.68)	1.49 (0.99–2.25)	**0.49 (0.29–0.82)**
PI's gender					
Female	1,661 (33.25)	190 (11.44)	126 (7.59)	0.99 (0.81–1.21)	**1.42 (1.10–1.83)**
Male	3,334 (66.75)	404 (12.12)	202 (6.06)	Ref.	Ref.
PI's age					
≤40	799 (16.76)	103 (12.89)	58 (7.26)	1.04 (0.81–1.33)	1.10 (0.80–1.51)
>40	3,969 (83.24)	459 (11.56)	253 (6.37)	Ref.	Ref.
Adequate experience of the research team					
Strongly agree/agree	4,310 (87.57)	562 (13.04)	213 (4.94)	**2.63 (1.73–4.00)**	**0.25 (0.19–0.32)**
Disagree/strongly disagree	612 (12.43)	28 (4.58)	111 (18.14)	Ref.	Ref.
Reviewer's gender					
Female	933 (18.84)	109 (11.68)	59 (6.32)	1.02 (0.80–1.29)	0.88 (0.64–1.22)
Male	4,018 (81.16)	484 (12.05)	268 (6.67)	Ref.	Ref.
World region of the reviewer					
Europe	2,838 (56.82)	322 (11.35)	208 (7.33)	Ref.	Ref.
North America	1,697 (33.97)	228 (13.44)	101 (5.95)	**0.62 (0.43–0.89)**	0.65 (0.37–1.15)
Other	460 (9.21)	44 (9.57)	19 (4.13)	0.84 (0.69–1.03)	1.27 (0.96–1.69)
Reviewer h-index					
≤15	2,279 (46.95)	292 (12.81)	164 (7.20)	1.08 (0.88–1.32)	1.09 (0.83–1.43)
>15	2,575 (53.05)	287 (11.15)	154 (5.59)	Ref.	Ref.

Note: Bold values indicate statistically significant findings.

Nagelkerke $R^2 = 0.102$.

CI = confidence interval; B = basic research; C = clinical research; E = epidemiological research; ref. = reference variable for RRR calculation.

In the design of our review process, significant efforts are made to ensure applicants are blinded in the first stage. All proposals are checked one by one to remove researcher's details, even though the research call instructions state that no information in the proposal may reveal a researcher's identity or affiliation. A random number of proposals are checked again with positive results in the majority of cases analysed.

We cannot ignore some points that favour concealing the identity of applicants and institutions. The main sources of bias in the peer-review process are related to the applicant, the project, or the potential relationship between the reviewer and the applicant (Lee 2000). In this last case, a strict policy of conflict of interest determines that the reviewer, either as collaborator or competitor, is not appropriate to evaluate the proposal, and this helps to prevent

Table 3. Description of the reasons for changing the qualitative assessment made by reviewers

Reasons to change	Classification	Type of change	Examples
Assessment of the team	Skills and experience	Positive	The experience of the researchers changed my mind regarding the unclear feasibility of the study. The study will be quite feasible. Even though some concerns regarding study design and data analysis remain, the study has potential. [ID 17]
		Negative	This project would be done by three people with rather varying degrees of expertise. Having reviewed the C's, it is doubtful to the reviewer whether this group may fulfil the ambitious goals set forth in the research plan. [ID 624]
	Multicentrism and collaboration	Positive	The different centres involved in the work for this proposal have been collaborating in the past with good results. This can guarantee the success of proposal, although the ambitions are very high. [ID 223]
		Negative	Given the multicentre structure, it is not sufficiently clear how this complicated effort will be coordinated. [ID 562]
	Composition and multidisciplinary	Positive	The multidisciplinary team is indeed composed of researchers having different backgrounds, expertise, and specialization. Most important, however, is the difference in scientific merits. [ID 667]
		Negative	My initial enthusiasm for this proposal is somewhat diminished by the lack of an expert in cell biology on the research team. The team is very strong in protein biochemistry, but not biological studies that are important for this project. [ID 1170]
Assessment of the proposal	Project	Positive	The applicant refers to important preliminary data that were not available in the first part of the evaluation; I have therefore reconsidered my evaluation in a positive way. [ID 2751]
	Project	Negative	Even though in the past this group has published studies on melanoma patients, in this project no clinical studies are scheduled. This makes the project not particularly interesting. [ID 939]
Suggestions about the team	Composition and multidisciplinary	Positive	The design and statistics of the protocol are flawed and suggest that a statistician would be a worthwhile member of the team. [ID 1592]
Suggestions about the proposal	Project	Positive	I advise the team of the project to focus on one or two tumours seen mostly in their centre, e.g. neuroblastoma (as most publications are on neuroblastoma and few are on sarcomas) and to focus only on one or two translational research issues (e.g. thyrosine hydroxylase . . .) and not to try to do many molecular techniques in the same project. [ID 1014]
Lack of information of the team	Skills and experience	Negative	The researchers did not follow instructions and provide information on their team. Instead they just provided a list of publications. They did describe a thesis which I presume that members of the team were mentors for these projects, but this was not clear. They also provided a list of funding, but it was not clear who received which funding. It is impossible to evaluate the researchers' expertise and experience fully with what they provided. [ID 2633]
Lack of information of the proposal	Project	Positive	I think this is a very interesting and important topic, but it is unclear how the researchers are going to disseminate the environmental information to patients, but of greater importance, there is no description on how any impact such information might have on asthma-related outcomes. Please see my initial review for further methodological issues. [ID 523]

ID = evaluation identification.

cronyism and nepotism. The bias related to a project usually comes from its innovative character or the degree to which it departs from mainstream science, the area where the majority of reviewers come from. However, the most common bias is related to the characteristics of researchers: gender, age, minority group and, specially, the well-known Matthew effect, related to the prestige, reputation, or recognition of researchers or institutions from which they come (Merton 1968). All these points assist us in favouring blinding.

4.1 Comparison with other studies

The only published study that assesses different ways of masking submissions, included in the Cochrane review, is a retrospective comparison between blind and open peer review of research proposals submitted to the Korea Science and Engineering Foundation (KOSEF). The process involved five reviewers for each proposal: three were sighted and two were blinded. A total of 1,978 proposals were sent to 917 reviewers; there were 562 answers, 331 from sighted reviewers and 231 from blinded ones. The study demonstrated that applicant's characteristics were the major factors leading to significant different evaluation scores between blinded and unblinded proposals. Results were considered proof of an obvious bias in the open evaluation of proposals towards researchers from top departments, senior researchers, and academically recognized researchers (Abdoul et al. 2012). In other words, it was reputation that made the difference. Similar results were found in the recent case of the NSF as mentioned above (Bhattacharjee 2012; Gurwitz et al. 2014), although no quantification was presented. Researcher and institutional prestige and even geographic location remain important sources of bias in research grant evaluation (Murray et al.

2016; Wahls 2016). The Matthew effect, or cumulative advantage, affects patterns of scientific collaboration, the growth of biological networks, the propagation of citations, scientific progress and impact, career longevity, as well as many other aspects of human culture (Perc 2014). In our case, it is worth mentioning that when the peer-review process is examined yearly, after the *ad hoc* committee (third and last stage) finishes its tasks, reviewers repeatedly praise the opportunity to assess a blinded proposal in the first stage.

In our study, the frequency of change in the second assessment and its direction (upward or downward) seemed to be affected by the year of the edition, the sum of the grant requested, the world region of the reviewer, and the gender of the PI. A probable explanation for why there were more changes in earlier editions than more recent ones might be that in earlier editions, improvement modifications in the review process were more common. For instance, modifications introduced included that only PIs should present a complete curriculum vitae with a short statement on the research team's experience since 2005, or the homogenization of the presentation of coordinated projects since 2007. Therefore, in more recent editions, the peer-review process has been more standardized, and only minor changes have been applied. We are unable to find another explanation for these differences. However, we must not forget that, as shown in different studies, there is always some inevitable degree of chance associated with the funding of grant applications and that a high number of reviewers is required to gain sufficient consistency to make decisions concerning proposals (Mayo et al. 2006).

Regarding differences depending on a reviewer's world region, it is worth mentioning that an evaluation of the Australian Research Council found that North American reviewers gave statistically significant higher ratings than those from other countries, such as those in Europe (Marsh et al. 2008). Also, the differences in the sum of the grant requested can be explained on the basis that lower amounts are related to single-group/institution proposals, to coordinated proposals having two groups from different institutions, and finally to proposals having three or more groups. This means that, in general, if lower numbers of PI/research teams are included in a proposal, a lower grant budget must be requested, and therefore, it will be more difficult to compensate for the expertise, specialization, and inclusion of different disciplines.

Finally, female PIs were associated with a negative change compared to no change and with respect to men. A meta-analysis of 66 different peer-review studies of grant applications showed a small gender effect in favour of men overall that was marginally statistically significant because of the large sample sizes, although a majority of individual studies showed no significant gender effect (Bornmann et al. 2007). Gender bias remains a disputed issue in the peer-review process with studies showing disparate results (Marsh et al. 2008; Van der Lee and Ellemers 2015). One recent review found only one study on interventions to mitigate gender bias in the peer review of grants (Tricco et al. 2017). This study found no difference in the proportion of women who were successful in receiving grant funding. Other authors speak of the 'Matilda effect', which highlights the historical tendency for women's work to be systematically omitted in the histories of scientific achievement (Rossiter 1993). However, we might assume that more recent editions are less gender biased and that some progress has been made in recent years (8.4% of the negative change in earlier editions versus 6.8% of the negative change in more recent editions, when the PI was female). An additional consideration is that fewer female PIs applied for these grants, though, in the end, the percentage of projects awarded grants to female PIs (33.1%) is almost exactly the same as their application ratio (33.7%). In any case, the design of our study precludes testing the role of gender experimentally, and association, as is well known, does not mean causality.

Although the evidence may not be conclusive, this study also shows that a change made in the second assessment is not affected by other characteristics that, according to the literature, might entail some bias such as a PI's age (Lee 2000) or the research area of a proposal, with some articles suggesting a bias against clinical research compared to molecular research (Marshall 1994). This fact is probably closely related to the reviewer's area of research, and thus, an accurate selection and matching of reviewers according to expertise (topic, and research area) has been attempted in the evaluation process.

4.2 Limitations of the study

The current study has certain limitations. First, the data come from a very specific peer-review process, where concealing applicants and institutions apply only to the first stage, so it might not be possible to extrapolate results to other systems. Secondly, the non-experimental nature of the study must be taken into account. We have no comparison group because data were extracted from the established peer-review process introduced in 2001 which, apart from the time constraints for the reviewers, is compensated for. We did not check if international reviewers were able to identify applicants in the blinded assessment phase, and it cannot be excluded in very specific research topics. However, with the help of electronic means, every effort is made to ensure anonymity when only the project is sent to reviewers (blinded assessment). The fact that the research peer-review process is an *ex ante* assessment of projects should be taken into account, and we do not know if this three-stage process with blinded applicants in the first stage improves the ex-post evaluation with a greater impact of the research, measured, for instance, with publications record and citations. Finally, since the questionnaire does not usually include a specific and compulsory direct question asking what the reasons for a change are, we only analysed an open field describing additional comments, so results were indirectly obtained.

4.3 Conclusions

In conclusion, blinding researchers/institutions in the first stage when assessing research grant proposals affects the second assessment in an average of 19% of the evaluations, from 10.5 to 27.7% depending on the year of edition. Attending to these rates, from 1 in 10 to 1 in 4, we believe that peer-review procedures that facilitate the focus only on the contents of a research proposal at first constitute a promising way to reduce some of the common biases described in the literature regarding researchers' characteristics and research grant decisions. Its implementation would reinforce transparency and accountability, so much in need nowadays for charities and public agencies financing research.

5. What is already known on this topic

To blind an applicant's identity or not has often been a topic of debate in the context of a peer-review process for research funds allocation.

Most of the evidence available on the effect of blinding on the quality of reviews comes from publications submitted to the editorial peer-review system.

6. What this study adds

To our knowledge, there are very few studies about the effect of blinding applicants in a sequential peer-review process of grant allocation for biomedical research projects.

Blinding applicants in the first stage helps to focus on a proposal and reduces biases related to a researcher's characteristics.

Acknowledgements

The authors would like to thank to Marta Aymerich, Maria Dolors Navarro, and Emília Sanchez for their involvement with the AQuAS peer-review process throughout its years in existence. The authors are also grateful to former technicians of AQuAS who contributed to its management.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Transparency

The lead authors (the manuscript's guarantors) affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

References

Abdoul, H. et al. (2012) 'Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices', *PLoS One*, 7/9: e46054.

Alam, M. et al. (2011) 'Blinded vs. Unblinded Peer Review of Manuscripts Submitted to a Dermatology Journal: A Randomized Multi-Rater Study', *British Journal of Dermatology*, 165/3: 563–7.

Baggs, J. G. et al. (2008) 'Blinding in Peer Review: The Preferences of Reviewers for Nursing Journals', *Journal of Advanced Nursing* 64/2: 131–8.

Bhattacharjee, Y. (2012) 'Science Funding. NSF's ′Big Pitch′ Tests Anonymized Grant Reviews', *Science*, 336/6084: 969–70.

Bornmann, L. R., Mutz, R., and Daniel, D. (2007) 'Gender Differences in Grant Peer Review: A Meta-analysis', *Journal of Informetrics*, 1/3: 226–38.

Budden, A. E. et al. (2008) 'Double-blind Review Favours Increased Representation of Female Authors', *Trends in Ecology and Evolution*, 23/1: 4–6.

Demicheli, V., and Di Pietrantonj, C. (2007) 'Peer Review for Improving the Quality of Grant Applications', *The Cochrane Database of Systematic Reviews*, 18/2: MR000003.

ESF (2011) *ESF Survey Analysis Report on Peer Review Practices* [Internet]. France: European Science Foundation <http://www.esf.org/fileadmin/Public_documents/. . ./pr_guide_survey.pdf'> accessed May 2015.

Garfield, E. (1987) 'Refereeing and Peer Review: Part 4. Research on the Peer Review of Grant Proposals and Suggestions for Improvement', *Essays of an Information Scientist*, 10: 27–33.

Gurwitz, D., Milanesi, E., and Koenig, T. (2014) 'Grant Application Review: The Case of Transparency', *PLoS Biology*, 12/12: e1002010.

Justice, A. C. et al. (1998) 'Does Masking Author Identity Improve Peer Review Quality? A Randomized Controlled Trial. PEER Investigators', *JAMA*, 280/3: 240–2.

Lee, M. (2000) 'The Bias of Sighted Reviewers in Research Proposal Evaluation: A Comparative Analysis of Blind and Open Review in Korea', *Scientometrics*, 48/1: 99–116.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008) 'Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability', *American Psychologist*, 63/3: 160–8.

Marshall, E. (1994) 'Does NIH Shortchange Clinicians?', *Science*, 265/5168: 20–1.

Mayo, N. E. et al. (2006) 'Peering at Peer Review Revealed High Degree of Chance Associated with Funding of Grant Applications', *Journal of Clinical Epidemiology*, 59/8: 842–8.

Merton, R. K. (1968) 'The Matthew Effect in Science: The Reward and Communication Systems of Science are Considered', *Science*', 159/3810: 56–63.

Mulligan, A., Hall, L., and Raphael, E. (2013) 'Peer Review in a Changing World: An International Study Measuring the Attitudes of Researchers', *Journal of the American Society for Information Science*, 64: 132–61.

Murray, D. L. et al. (2016) 'Bias in Research Grant Evaluation has Dire Consequences for Small Universities', *PLoS One*, 11/6: e0155876.

Perc, M. (2014) 'The Matthew Effect in Empirical Data', *Journal of the Royal Society, Interface*, 11/98: 20140378.

Regehr, G., and Bordage, G. (2006) 'To Blind or not to Blind? What Authors and Reviewers Prefer', *Medical Education*, 40/9: 832–9.

Rossiter, M. W. (1993) 'The Matthew Matilda Effect in Science', *Social Studies of Science*, 23/2: 325–41.

Tabak, L. A., and Collins, F. S. (2011) ' Sociology. Weaving a Richer Tapestry in Biomedical Science', *Science*, 333/6045: 940–1.

Tricco, A. C. et al. (2017) 'Strategies to Prevent or Reduce Gender Bias in Peer Review of Research Grants: A Rapid Scoping Review', *PLoS One*, 6, 12/1: e0169718.

Triggle, C. R., and Triggle, D. J. (2007) 'What is the Future of Peer Review? Why is there Fraud in Science? Is Plagiarism out of Control? Why do Scientists do Bad Things? Is it all a Case of: "all that is Necessary for the Triumph of Evil is that Good Men do Nothing"?', *Vascular Health and Risk Management*, 3: 39–53.

Van der Lee, R., and Ellemers, N. (2015) 'Gender Contributes to Personal Research Funding Success in The Netherlands', *Proceedings of the National Academy of Sciences of the USA* , 112/40: 12349–53.

Van Rooyen, S. et al. (1998) 'Effect of Blinding and Unmasking on the Quality of Peer Review: A Randomized Trial', *JAMA*, 280/3: 234–7.

Van Rooyen, S. et al. (1999) 'Effect of Blinding and Unmasking on the Quality of Peer Review', *Journal of General Internal Medicine*, 14/10: 622–4.

Wahls, W. P. (2016) 'Biases in Grant Proposal Success Rates, Funding Rates and Award Sizes Affect the Geographical Distribution of Funding for Biomedical Research', *PeerJ*, 4: e1917.

Wessely, S. (1998) 'Peer Review of Grant Applications: What do we Know?', *Lancet*, 352/9124: 301–5.

ANNEX 2. PUBLICACIÓ #2 (AVALUACIÓ EX-POST)

Research Evaluation, 28(3), 2019, 253–262
doi: 10.1093/reseval/rvz007
Advance Access Publication Date: 16 April 2019
Article

OXFORD

Development and validation of a questionnaire to measure research impact

Maite Solans-Domènech[1,2,3,*], Joan MV Pons[1,2], Paula Adam[1,2], Josep Grau[4] and Marta Aymerich[3,5]

[1]Agency for Health Quality and Assessment of Catalonia (AQuAS), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [2]Epidemiology and Public Health Network (CIBER ESP), Roc Boronat 81-95, Barcelona, Catalonia 08020, Spain, [3]Faculty of Health Sciences, Universitat Oberta de Catalunya (UOC), Rambla del Poblenou, 156, Barcelona, Catalonia 08018, Spain, [4]Research Planning Unit, Universitat Oberta de Catalunya (UOC), Rambla del Poblenou, 156, Barcelona, Catalonia 08018, Spain, and [5]eHealth Center (eHC), Universitat Oberta de Catalunya (UOC), Av. Tibidabo, 39-43, Barcelona, Catalonia 08035, Spain

*Corresponding author. Email: mtsolans@gencat.cat

Abstract

Although questionnaires are widely used in research impact assessment, their metric properties are not well known. Our aim is to test the internal consistency and content validity of an instrument designed to measure the perceived impacts of a wide range of research projects. To do so, we designed a questionnaire to be completed by principal investigators in a variety of disciplines (arts and humanities, social sciences, health sciences, and information and communication technologies). The impacts perceived and their associated characteristics were also assessed. This easy-to-use questionnaire demonstrated good internal consistency and acceptable content validity. However, its metric properties were more powerful in areas such as knowledge production, capacity building and informing policy and practice, in which the researchers had a degree of control and influence. In general, the research projects represented an stimulus for the production of knowledge and the development of research skills. Behavioural aspects such as engagement with potential users or mission-oriented projects (targeted to practical applications) were associated with higher social benefits. Considering the difficulties in assessing a wide array of research topics, and potential differences in the understanding of the concept of 'research impact', an analysis of the context can help to focus on research needs. Analyzing the metric properties of questionnaires can open up new possibilities for validating instruments used to measure research impact. Further to the methodological utility of the current exercise, we see a practical applicability to specific contexts where multiple discipline research impact is requires.

Key words: surveys and questionnaires; research impact assessment; validation study

Introduction

Over the past three decades, increasing attention has been paid to the social role and impact of research carried out at universities. National research evaluation systems, such as the UK's Research Excellence Framework (REF) (Higher Education Funding Council of England et al. 2015) and the Excellence in Research for Australia (Australian Research Council 2016) are examples of assessment tools that address these concerns. These systems identify and define how research funding is allocated based on a number of dimensions of the research process, including impact of research. (Berlemann and Haucap 2015).

Being explicit about the objective of the impact assessment is emphasized in the International School on Research Impact Assessment (ISRIA) statement (Adam et al. 2018) a 10-point guideline for an effective research impact assessment that includes four purposes: advocacy, analysis, allocation, and accountability. The last one emphasizes transparency, efficiency, value to the public and a return for the investment. With mounting concern about the

relevance of research outcomes, funding organizations are increasingly expecting researchers to demonstrate that investments result in tangible improvements for society (Hanney et al. 2004). This accountability is intended to ensure resources have been appropriately utilized and is strongly linked to the drive for value-for-money within health services and research (Panel on the return on investments in health research 2009). As policy-makers and society expect science to meet societal needs, scientists have to prioritize social impact, or risk losing public support (Poppy 2015).

To meet these expectations, the Universitat Oberta de Catalunya (UOC) has embraced a number of pioneering initiatives in its current Strategic Plan, which includes the promotion of Open Knowledge, a specific measure related to the social impact of research, (Universitat Oberta de Catalunya 2017) and the development of an institution wide action plan to incorporate it in research evaluation. The UOC is currently investigating how to implement the principals of the DORA Declaration in institutional evaluation processes, taking into account 'a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice' ('San Francisco Declaration on Research Assessment (DORA)' n.d.). The UOC is also taking the lead in meeting the Sustainable Development Goals (SDG) of the UN 2030 Agenda,(Jørgensen and Claeys-Kulik 2018) having been selected by the International Association of Universities as one of the 16 university cluster leaders around the world to lead the SDGs ('IAU HESD Cluster | HESD - Higher Education for Sustainable Development portal' n.d.).

The term 'research impact' has many definitions. On a basic level, the 'academic impact' is understood as benefits for further research, while 'wider and societal impact' includes the outcomes that reach beyond academia. In our study we will include both categories and refer to 'research impact' as any type of output or outcome of research activities that can be considered a 'positive return or payback' for a wide range of beneficiaries, including people, organizations, communities, regions, or other entities. The pathways linking science, practice, and outcomes are multifaceted and complex (Molas-Gallart et al. 2016). Indeed, the path from new knowledge to its practical application is neither linear nor simple; the stages may vary considerably in terms of duration, and many impacts of research may not be easily measurable or attributable to a concrete result of research (Figure 1). This outputs and outcomes generated by research characteristics (inputs and processes) are context dependant (Pawson 2013). Therefore, a focus on process is fundamental to understanding the generation of impact.

Surveys are among the most widely used tools in research impact evaluation. Quantitative approaches as surveys are suggested for accountability purposes, as the most appropriate way that calls for transparency (Guthrie et al. 2013). They provide a broad overview of the status of a body of research and supply comparable, easy-to-analyze data referring to a range of researchers and/or grants. Standardization of the approach enhances this comparability and minimizes researcher bias and subjectivity, particularly in the case of web or postal surveys. Careful wording and question construction increases the reliability of resulting data (Guthrie et al. 2013). However, while ex-ante assessments instruments for research proposals have undergone significant study, (Fogelholm et al. 2012; Van den Broucke et al. 2012) the metric properties of research evaluation instruments have received little attention (Aymerich et al. 2012). 'Internal consistency' is generally considered evidence of internal structure, (Clark and Watson 1995) while the measurement of 'content validity' attempts to demonstrate that the elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose (Nunnally and Bernstein 1994).

As the demand for monitoring research impact increases across the world, so does the need for research impact measures that demonstrate validity. Therefore, the aim of this study is to develop and test the internal consistency and the content validity of an instrument designed for accountability purposes to measure the perceived impacts of a wide range of competitively funded research projects, according to the perspectives of the principal investigators (PIs). The study will also focus on the perceived impacts and their characteristics.

Methods

A cross-sectional survey was used to assess the research undertaken at UOC. This research originates from four knowledge areas: arts and humanities, social sciences, health sciences, and information and communication technologies (ICT). Research topics include 'identity, culture, art and society'; 'technology and social action'; 'globalization, legal pluralism and human rights'; 'taxation, labour relations and social benefits'; 'internet, digital technologies and media'; 'management, systems and services in information and communications'; and 'eHealth'. UOC's Ethics Committee approved this study.

Study population

The study population included all PIs with at least one competitively funded project (either public or private) at local, regional, national, or international level completed by 2017 (n = 159).

The questionnaire

An on-line questionnaire was designed for completion by project PIs in order to retrospectively determine the impacts directly attributed to the projects. The questions were prepared based on the team's prior experience and questionnaires published in scientific literature. (Wooding et al. 2010; Hanney et al. 2013) The questionnaire was structured around the multidimensional categorization of impacts in the Payback Framework. (Hanney et al. 2017)

The Payback Framework has been extensively tested and used to analyze the impact of research in various disciplines. It has three elements: first, a logic model which identifies the multiple elements that form part of the research process and contribute to achieving impact; second, two 'interfaces', one referring to the project specification and selection, the other referring to the dissemination of research results; and third, a consideration of five impact categories: *knowledge production* (represented by scientific publications or dissemination to non-scientific audiences); *research capacity building* (research training, new collaborations, the securing of additional funding or improvement of infrastructures); *informing policy and product development* (research used to inform policymaking in a wide range of circumstances); *social benefits* (application of the research within the discipline and topic sector); and *broader economic benefits* (commercial exploitation or employment) (Hanney et al. 2013).

Our instrument included four sections. The first section recorded information on the PIs, including their sex, age, and the number of years they had been involved in research. The second focused on the nature of the project itself (or a body of work based on

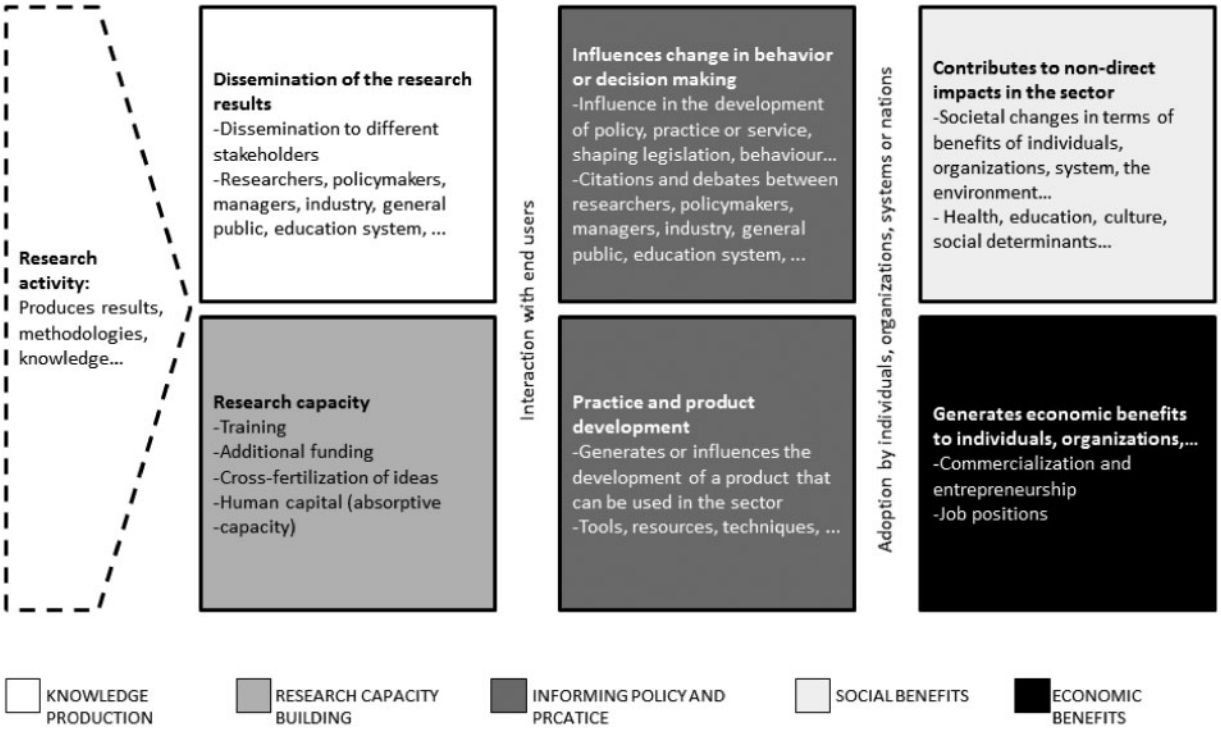


Figure 1. Effects of research impact.

continuation/research progression projects). PIs involved in more than one project (or a set of projects within the same body of work) were instructed to select one, in order to reduce the time needed to complete the survey and thereby to increase response rate. This section included the discipline, the main topic of research, the original research drivers, interaction with potential users of the research during the research processes, and funding bodies. The third section addressed the PIs' perceptions of the impact of the research project, and was structured around the five impact categories of the aforementioned Payback Framework. The last section included general questions, one of which sought to capture other relevant impacts that might not fall within one of the previous five categories. The final question requested an evaluation (as a percentage) of the contribution/attribution of the research to the five impact categories. Respondents were required to decide the level of the contribution/attribution of the impacts according to three answer categories: limited, contribution from 1 to 30%; moderate, contribution from 40 to 60%; and significant, contribution from 70 to 100%).

Questionnaire items included questions with dichotomous answers (yes/no) and additional open box questions for a brief descriptions of the impacts perceived.

Prior to testing, we reviewed the abstracts of 72 REF2014 impact case studies (two per knowledge area). REF2014 (Higher Education Funding Council of England et al. 2015) is the first country-wide exercise to assess the impact of university research beyond academia and has a publicly available database of over 6,000 impact case studies, grouped in 34 subject-based units of assessment. Case studies were randomly selected and the impacts found in each mapped onto the most appropriate items and dimensions of the questionnaire. This review helped to reformulate and add questions, especially in the sections on *informing policy and practice* and *social benefits*.

Data collection

The questionnaire was sent to experts in various disciplines with a request for feedback on the relevance of each item to the questionnaire's aim (impact assessment), which they rated on a 4-point scale (0 = 'not relevant', 1 = 'slightly relevant', 2 = 'quite relevant', 3 = 'very relevant') according to the definition of research impact included in our study (defined above). The experts were also asked to evaluate whether the items covered the important aspects or whether certain components were missing. They could also add comments on any item.

The PIs were contacted by email. They were informed of the objectives of the study and assured that the data would be treated confidentially. They received two reminders, also by email.

Analysis

A quality control exercise was performed prior to data analysis. The data were processed and the correct classification of the various impacts checked by comparing the yes/no responses with the information provided in the additional open box questions. No alterations were required after these comparisons. Questionnaire results provided a measure of the number of research projects contributing to a particular type of impact; therefore, to estimate each level of impact we calculated the frequency of its occurrence in relation to the number of projects. A Chi-squared test was used to test for group differences.

Internal consistency was assessed by focusing on the inter-item correlations within the questionnaire, indicating how well the items fitted together theoretically. This was performed using Cronbach's alpha (α). An alpha between 0.70 and 0.95 was considered acceptable (Peterson 1994).

An expert opinion index was used to estimate content validity at the item level. This index was calculated by dividing the number of

experts providing a score of 2 or 3 by the total number of answers. Due to the diverse array of disciplines and topics under examination, values were calculated for all experts and for the experts of each discipline. These were considered acceptable if the level of endorsement was >0.5.

All data were introduced into the statistical programme SPSS18, and the level of significance set at 0.05 for all tests.

Results

Sixty-eight PIs answered the questionnaire, a response rate of 42.8%. Respondents took an average of 26 minutes to complete the questionnaire. Table 1 shows the sample characteristics. Significant differences were found between the respondents and non-respondents for knowledge area ($p = 0.014$) and age group ($p = 0.047$). Arts and humanities investigators and PIs older than 50 years were more frequent among non-respondents. The proportion of women did not differ significantly between respondents and non-respondents ($p = 0.083$).

Impact and its characteristics

An impact on *knowledge production* was observed in 97.1% of the projects, and an impact on *capacity building* in 95.6%. Lower figures were recorded for *informing policy and practice* (64.7%), and lower still for *economic benefits* (33.8%), and for *social benefits* (32.4%), although results were based on a formal evaluation in only 11, 8% of the cases included in *social benefits*. Estimations of the contribution of projects to the different impact levels were considered significant (between 70% and 100%) to *knowledge production*, moderate (between 40% and 60%), to *capacity building*, and

Table 1. Sample characteristics

		Respondents, n = 68 (42.8%)	Non-respondents, n = 91 (57.2%)
Knowledge area*	Social Sciences	42 (61.8)	47 (51.6)
	Information and Communication Technologies	14 (20.6)	16 (17.6)
	Health Sciences	7 (10.3)	7 (4.4)
	Arts and Humanities	5 (7.4)	24 (26.4)
Research subject[a]	Education	25 (36.8)	–
	Internet, digital technologies and the media	20 (29.4)	–
	Computation and artificial intelligence	15 (22.1)	–
	Health and sustainable lifestyles	14 (20.6)	–
	Art, culture and identity	11 (16.2)	–
	Society, social action and the environment	10 (14.7)	–
	Governance and social movements	7 (10.3)	–
	Management, systems and services in information and communications and innovation	4 (5.9)	–
	Globalization, legal pluralism and human rights	4 (5.9)	–
	Language, literature and cognition	4 (5.9)	–
	Tourism	2 (2.9)	–
	Others	8 (11.8)	–
Original impetus for the project[a]	Scientific curiosity	32 (47.1)	–
	The need to fill certain gaps in knowledge	38 (55.9)	–
	Targeting to a practical application	39 (57.4)	–
	Personal professional experience	23 (33.8)	–
	Commissioned by third parties	2 (2.9)	–
Time elapse since the beginning of the project	Less than 4 years	17 (24.5)	–
	4–9 years	35 (50.7)	–
	More than 9 years	16 (23.9)	–
	Unknown	1 (1.4)	–
Interaction with end users[a]	Before the research process	24 (35.3)	–
	During the research process	48 (70.6)	–
	After the research process	41 (60.3)	–
	No interaction	7 (10.3)	–
PI's gender	Woman	34 (50.0)	33 (36.3)
	Man	34 (50.0)	58 (63.7)
PI's age (years)*	<30	3 (4.4)	2 (2.2)
	31–40	17 (25.0)	12 (13.2)
	41–50	38 (55.9)	45 (49.5)
	>50	10 (14.7)	28 (30.8)
	Unknown	–	4 (4.4)
PI's research experience	<5 years	5 (7.3)	–
	6–10 years	13 (19.1)	–
	>10 years	50 (73.5)	–

[a]Answers could include more than one response. PI: principal investigator.

*$p \leq 0.05$.

limited (1–30%) to *informing policy and practice*, *social benefits* and *economic benefits*. No additional impacts were reported.

Figure 2 shows the different impact categories and the distribution of impact subcategories. The size of the bars indicates the percentage of projects in which this specific impact occurred, according to the PIs.

Statistically significant differences were found according to the original impetus for the project: for projects intended to fill certain gaps in knowledge, the greatest impact was observed in *knowledge production* ($p = 0.01$) and *capacity building* ($p = 0.03$), while for projects targeting to a practical application, the greatest impact was observed in *informing policy and practice* ($p = 0.05$) and in *social*

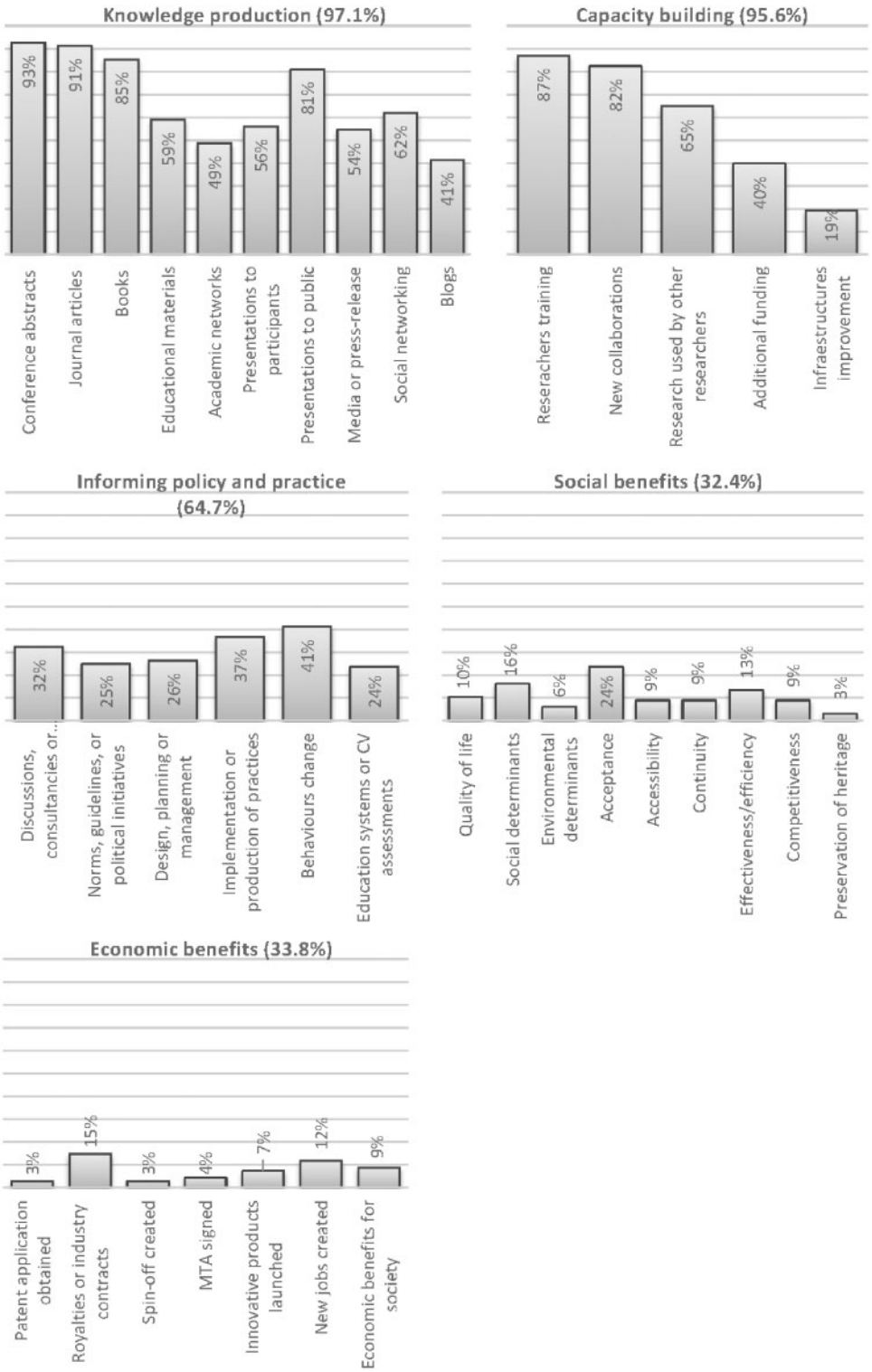


Figure 2. Achieved impact bars, according to level (n = 68).

Table 2. Internal consistency for each domain (impact level)

Domain	Cronbach's alpha
Knowledge Production	0.74
Capacity Building	0.74
Informing Policy and Practice	0.82
Social Benefits	0.89
Economic Benefits	0.47
Total domains	0.89

benefits ($p = 0.01$). In general, projects that interacted with end users had more impact in the levels of *knowledge production* ($p = 0.01$), *capacity building* ($p = 0.03$), and *social benefits* ($p = 0.05$). Projects that had begun over four years before the survey was completed was correlated with *knowledge production* ($p = 0$, 04), and PIs over 40 years of age and those with over 3 years research experience were correlated with more frequent impacts on *knowledge production* and *capacity building* ($p \leq 0.01$). No differences were found regarding the gender of PI's. The size of the differences can be found in the Supplementary Table S1.

Internal consistency and content validity

The Cronbach's alpha score, which measures the internal consistency of the questions, was satisfactory ($\alpha = 0.89$). Table 2 shows its value in each domain (impact level). Internal consistency was satisfactory in all domains with the exception of *economic benefits*. However, the removal of any of the questions would have resulted in an equal or lower Cronbach's alpha.

Thirteen of the 17 experts contacted completed the content validity form and assessed whether the content of the questionnaire was appropriate and relevant to the purpose of the study. Seven were from social sciences and humanities, four from health sciences and two from ICT; 39% were women. All had longstanding experience as either researchers or research managers. The experts scored the 45 items according to their relevance and 76% of the ratings ($n = 34$) had an index of 0.5 or greater. The results for each item are shown in Table 3. In accordance with the expert review, an item relating to 'new academic networks' was added.

Ninety-one percent of the items in *knowledge production* were rated acceptable (expert opinion index ≥ 0.5), as were 89% of the items in *capacity building*, 83% of the items in *informing policy and practice*, and 63% of the items in *social benefits*. In contrast, only 43% of the items (three out of seven) in the *economic benefits* domain achieved an acceptable rating. Some items were of higher relevance in specific fields: for example, items relating to health and social determinants were considered acceptable by health experts; training for final undergraduate's projects was considered acceptable by ICT experts; influencing education systems and curricular assessments, was considered acceptable by social sciences and humanities, and ICT experts; and commercialization items were considered acceptable by health and ICT experts (Table 3).

Discussion

In this study, we tested the metric properties of a questionnaire designed to record the impact of university research originating from various disciplines. Tests of this kind, although rare in research impact assessment, are common in other study areas such as patient-reported outcome measures, education and psychology. The questionnaire displayed good internal consistency and acceptable content validity in our context. Internal consistency for all items on the instrument was excellent demonstrating that they all measured the same construct. However, since 'impact' is a multidimensional concept and, by definition, Cronbach's alpha 'indicates the correlation among items that measure one single construct', (Osburn 2000) the internal consistency of each of the five domains required evaluation; this was found to be excellent in all cases except *economic benefits*. Low internal consistency in this domain may be related to the fact it contained relatively few items, and/or the fact that most of the researchers who answered the questionnaire worked in the social sciences and humanities, and therefore impacts relating to transfer, commercialization and innovation were less likely to occur. An alternative possibility is that the items are, in fact, measuring more than one construct.

There is a consensus in the literature that content validity is largely a matter of judgment, (Mastaglia et al. 2003) as content validity is not a property of the instrument, but of the instrument's interpretation. We therefore incorporated two distinct phases in our study. In the first phase of development, conceptualization was enhanced through the analysis and mapping of the impacts of the randomly selected REF case; in the second the relevance of the scale's content was evaluated through expert assessment. The expert assessment revealed that some items did not achieve acceptable content validity, especially in the domains of *social benefits* and *economic benefits*. However, it should be taken into account that while many of the items in the questionnaire were generic and thus relevant for all fields, a number were primarily specific to one field, and therefore, more relevant for experts in a particular field. Content validity was stronger in the domains 'closest' to the investigators. This may be due to the most frequently recognized impacts being both in areas where researchers have a degree of control and influence, (Kalucy et al. 2009) and those which have been 'traditionally' used to measure research. In other words, their understanding of the concept of impact in the *knowledge production*, *capacity building* and *informing policy and practice* domains: that is, those at the intermediate level (secondary outputs) display greater homogeneity (Kalucy et al. 2009).

Use of an online questionnaire in this research impact study provided data on a wide range of benefits deriving from UOC's funded projects at a particular moment and its results address a message of accountability. Questionnaires can provide insights into respondents' viewpoints and can systematically enhance accountability. Although assuming that PIs will provide truthful responses about the impact of their research is clearly a potential limitation, Hanney et al (2013) demonstrate that researchers do not routinely exaggerate the impacts of their research, at least in studies like this one, where there is no clear link between the replies given and future funding. International guidelines on research impact assessment studies recommend the use of a combination of methods to achieve comprehensive, robust results. (Adam et al. 2018) However, the primary focus of this study was the quality and value of the survey instrument itself, therefore the issue of triangulating the findings with other methods was not explored. The questionnaire could be applied in future studies to select projects that require a more in-depth and closer analysis, such as how an understanding of scientific processes works in this context. Previous attempts have been made to assess the impact of university's research in our context, but these have been restricted to the level of outputs (i.e. publications and patents), (Associació Catalana d'Universitats Públiques (ACUP) 2017) or inputs' level (i.e. contributions to Catalan GDP) (Suriñach et al. 2017).

Table 3. Content validity of items according to experts (n = 13)

Domain	Item	Number of answers	All experts (n = 13)	Social experts (n = 7)	Health experts (n = 4)	ICT experts (n = 2)
Knowledge Production	Presenting research findings in abstracts	13	0.77*	0.71*	0.75*	1.00*
	Presenting research findings in journal articles	13	0.85*	0.71*	1.00*	1.00*
	Presenting research findings in books or book chapters	13	0.69*	0.71*	0.50*	1.00*
	Presenting the research findings in educational materials	13	0.62*	0.57*	0.50*	1.00*
	Presentations of research findings to the public/patients/end-users	13	0.62*	0.43	1.00*	0.50*
	Presentations to the project volunteers	13	0.69*	0.57*	0.75*	1.00*
	Been mentioned by the media or the subject of a press release/conference	13	0.85*	0.71*	1.00*	1.00*
	Published through social networks	13	0.62*	0.43	1.00*	0.50*
	Published in influential blogging sites	12	0.67*	0.57*	1.00*	0.50*
	Concerts, recordings, or music hall presentations	9	0.11	0.25	0.00	0.00
Capacity Building	Training for PhD students	13	0.92*	0.86*	1.00*	1.00*
	Training for master's degree students	13	0.69*	0.57*	0.75*	1.00*
	Training for final undergraduate's projects	13	0.31	0.43	0.00	0.50*
	New collaborations at national level	11	0.91*	0.83*	1.00*	1.00*
	New collaborations at international level	12	0.92*	0.83*	1.00*	1.00*
	New academic networks	8	0.75*	0.75*	0.50*	1.00*
	Additional funding to create new research projects	13	0.77*	0.57*	1.00*	1.00*
	Additional funding for the research group	13	0.92*	0.86*	1.00*	1.00*
	Research or methods used by other researchers in subsequent research	9	0.78*	0.67*	0.75*	1.00*
	Contribution to the improvement of research infrastructures	13	0.62*	0.57*	0.50*	1.00*
Informing Policy and Practice	Informing as evidence in debates, discussions, or consultancies	13	0.62*	0.71*	0.50*	0.50*
	Informing as evidence in the formulation of norms, guidelines, political initiatives or recommendations by government bodies or other regulators	12	0.58*	0.67*	0.50*	0.50*
	Contribution in the design, planning and management of services and priorities	12	0.50*	0.50*	0.50*	0.50*
	In the implementation, adoption or production of practices within and beyond the professional world	12	0.58*	0.50*	0.50*	1.00*
	Influencing the behaviour of professionals or other people	13	0.54*	0.57*	0.50*	0.50*
	Influencing education systems and curricular assessments	10	0.40	0.50*	0.25	0.50*
Social Benefits	Improving health	11	0.27	0.20	0.50*	0.00
	Improving quality of life	11	0.55*	0.60*	0.50*	0.50*
	Improving social and cultural determinants	11	0.64*	0.80*	0.50*	0.50*
	Improving environmental determinants	11	0.36	0.40	0.50*	0.00
	Improving acceptability	11	0.73*	0.80*	0.75*	0.50*
	Improving accessibility*	11	0.55*	0.60*	0.50*	0.50*
	Improving continuity	9	0.56*	0.67*	0.50*	0.50*
	Improving effectiveness or efficiency	10	0.50*	0.50*	0.50*	0.50*
	Improving safety	10	0.20	0.25	0.25	0.00
	Improving well-being and social benefits	11	0.64*	0.80*	0.50*	0.50*
	Improving heritage preservation	10	0.20	0.25	0.25	0.00
	Improving competitiveness and development of stimuli	10	0.50*	0.50*	0.25	1.00*

(continued)

Table 3. Continued

Domain	Item	Number of answers	All experts (n = 13)	Social experts (n = 7)	Health experts (n = 4)	ICT experts (n = 2)
Economic Benefits	Patent application obtained	13	0.23	0.00	0.50*	0.50*
	Generating revenue from royalties, equities, industry contracts or any other compensation	13	0.23	0.14	0.50*	0.00
	Leading to the creation of a new business spin-off or start-up company	13	0.31	0.14	0.50*	0.50*
	Leading to Material Transfer Agreements	13	0.23	0.00	0.50*	0.50*
	Bringing innovations, products or devices to market	9	0.56*	0.33	0.50*	1.00*
	Creation of new jobs	13	0.62*	0.57*	0.75*	0.50*
	Bringing wider economic impacts	13	0.54*	0.57*	0.50*	0.50*

*Items rated greater than or equal to 0.5; ICT: information and communication technologies.

Evaluated as a whole, the research projects covered in this study was effective in the production of knowledge and the development of research skills in individuals and teams. This funded research has helped to generate new knowledge for other researchers and, to a lesser extent, for non-academic audiences. It has consolidated the position of UOC researchers (both experienced and novice) within national and international scientific communities, enabling them to develop and enhance ability to conduct quality research (Trostle 1992).

Assessing the possible wider benefits of the research process (in terms of *informing policy and practice*, *social benefits* and *economic benefits* for society) proved more problematic. The relatively short period that had elapsed since the projects finished might have limited the assessment of impact. There was a striking disparity, in our results, between the return on research measured in terms of scientific impact (*knowledge production* and *capacity building*), notably high and uniform, and the limited and uneven contribution to wider benefits. This disparity is not a local phenomenon, but a recurrent finding in contemporary biomedical research worldwide. The Retrosight study, (Wooding et al. 2014), which analyzed cardiovascular and stroke research in the United Kingdom found no correlation between knowledge production and the broader social impact of research. Behavioural aspects such as researcher engagement with potential users of the research or mission-oriented projects (targeted to practical applications) were associated with higher social benefits. This might be interpreted as strategic thinking on the part of researchers, in the sense that they consider the potential 'mechanisms' that might enhance the impact of their work. These results do not appear to be exceptional, since the final impact of research is influenced by the extent to which the knowledge obtained is made available to those in a position to use it.

Although the response rate was lower than expected, 43% is within the normal range for on-line surveys. (Shih and Xitao 2008) In addition, arts and humanities researchers were underrepresented among PIs, but not between experts for considering content validity. One possible reason for this is that investigators are not fully aware of the influence of their research; another is the belief that research impact assessment studies are unable to provide valuable data about how arts and humanities research generates value. (Molas-Gallart 2015) Arts and humanities is a discipline where in some cases the final objective of the research is not a practical application, but rather to change behaviours or people perspectives, which are therefore, more difficult to measure. According to Ochsner et al. (2012) there is a missing link between indicators and humanities scholars' notions of quality. However, questionnaires have been used to successfully measure the impact of arts and humanities research, including in an approach adapted from the Payback Framework (Levitt et al. 2010), and research impact analyses such as REF2014 (Higher Education Funding Council of England et al. 2015) and the special issue of Arts and Humanities in Higher Education on the public value of arts and humanities research (Benneworth 2015) have demonstrated that research in these disciplines may have many implications for society. Research results provide guidance and expertise and can be easily transferred to public debates, policies and institutional learning.

Weiss describes the rationale and conceptualization of assessment activities relating to the social impact of research as an open challenge (Weiss 2007). As well as the well-known practice of attributing impact to a sole research project and the time-lag between the start of a research project and the attainment of a specific impact, in this study we also had the challenge to assess the impact of research from a diverse variety of topics and disciplines. Research impact studies are prevalent in disciplines such as health sciences (Hanney et al. 2017) and agricultural research (Weißhuhn et al. 2018) but less common in the social sciences and humanities, despite the REF2014 results revealing a wide array of impacts associated with various disciplines. (Higher Education Funding Council of England et al. 2015) Our challenge was to analyze projects from highly diverse disciplines—social sciences, humanities, health sciences, and ICTs—and assess their varied impacts on society. We have attempted to develop a flexible and adaptable approach to assessing research impacts by utilizing a diverse amalgamation of indicators, including impact subcategories. However, due to 'cultural' differences between disciplines, we cannot guarantee that PIs from different knowledge areas have a homogeneous understanding of 'research impact': indeed the diversity of respondents when assessing the relevance of questionnaire items suggests otherwise. For this reason, a context analysis in which research is carried out and assessed, as described in the literature (Adam et al. 2018) may help to decide which questionnaire items or domains should be included or removed in future studies.

To conclude, this study demonstrates that the easy-to-use questionnaire developed here is capable of measuring a wide range of research impact benefits and provides good internal consistency. Analyzing the metric properties of instruments used to measure research impact and establishing their validity will significantly

contribute to research impact assessment and stimulate and extend reflection on the definition of research impact. Therefore, this questionnaire can be a powerful instrument to measure research impact when considered in context. The power of this instrument will be significantly improved when combined with other methodologies.

What is already known about this topic

Surveys are widely used in research impact evaluation. They provide a broad overview of the state of a body of research, and supply comparable, easily analyzable data referring to a range of researchers and/or grants. The standardization of the approach enhances this comparability.

What this study adds

To our knowledge, the metric properties of impact assessment questionnaires have not been studied to date. The analysis of these properties can determine the internal consistency and content validity of these instruments and the extent to which they measure what they are intended to measure.

Supplementary data

Supplementary data is available at *Research Evaluation Journal* online.

Acknowledgements

We thank the UOC principal investigators for providing us with their responses.

Funding

This project did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Transparency

The lead authors (the manuscript's guarantors) affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

References

Adam, P. et al. (2018) 'ISRIA Statement: Ten-Point Guidelines for an Effective Process of Research Impact Assessment', *Health Research Policy and Systems*, 16/1, DOI: 10.1186/s12961-018-0281-5.

Associació Catalana d'Universitats Públiques (ACUP) (2017) *Research and Innovation Indicators of Catalan Public Universities*. Report 2016.

Australian Research Council (2016) *State of Australian University Research 2015-2016: Volume 1 ERA National Report*. Canberra: Commonwealth of Australia.

Aymerich, M. et al. (2012) 'Measuring the Payback of Research Activities: A Feasible Ex-Post Evaluation Methodology in Epidemiology and Public Health', *Social Science and Medicine*, 75/3: 505–10.

Benneworth, P. (2015) 'Putting Impact into Context: The Janus Face of the Public Value of Arts and Humanities Research', *Arts and Humanities in Higher Education*, 14/1: 3–8.

Berlemann, M. and Haucap, J. (2015) 'Which Factors Drive the Decision to Opt out of Individual Research Rankings? An Empirical Study of Academic Resistance to Change', *Research Policy*, 44/5: 1108–15.

Clark, L. A. and Watson, D. (1995) 'Constructing Validity: Basic Issues in Objective Scale Development', *Psychological Assessment*, 7/3: 309–19.

Fogelholm, M. et al. (2012) 'Panel Discussion Does Not Improve Reliability of Peer Review for Medical Research Grant Proposals', *Journal of Clinical Epidemiology*, 65/1: 47–52.

Guthrie, S. et al. (2013) *Measuring Research: A Guide to Research Evaluation Frameworks and Tools*. RAND Corporation.

Hanney, S. et al. (2004) 'Proposed Methods for Reviewing the Outcomes of Health Research: The Impact of Funding by the UK's "Arthritis Research Campaign"', *Health Research Policy and Systems*, 2/1: 4.

Hanney, S. et al. (2013) 'Conducting Retrospective Impact Analysis to Inform a Medical Research Charity's Funding Strategies: The Case of Asthma UK', *Allergy, Asthma, and Clinical Immunology: Official Journal of the Canadian Society of Allergy and Clinical Immunology*, 9/1: 17.

Hanney, S. et al. (2017) 'The Impact on Healthcare, Policy and Practice from 36 Multi-Project Research Programmes: Findings from Two Reviews', *Health Res Policy Syst*, 15/1: 26.

Higher Education Funding Council of England, et al. (2015) *The Nature, Scale and Beneficiaries of Research Impact: An Initial Analysis of Research Excellence Framework (REF) 2014 Impact Case Studies*. London: HEFCE.

'IAU HESD Cluster | HESD - Higher Education for Sustainable Development portal' (n.d.) <http://iau-hesd.net/en/contenu/4648-iau-hesd-cluster.html> accessed 17 Dec 2018.

Jørgensen, T. E. and Claeys-Kulik, A.-L. (2018) *Universities' Strategies and Approaches Towards Diversity, Equity and Inclusion. Examples from Across Europe*. Brussels: European University Association.

Kalucy, E. C. et al. (2009) 'The Feasibility of Determining the Impact of Primary Health Care Research Projects Using the Payback Framework', *Health Research Policy and Systems*, 7: 11.

Levitt, R., Celia, C. and Diepeveen, S. (2010) *Assessing the Impact of Arts and Humanities Research at the University of Cambridge*. Technical Report. RAND Corporation, 104.

Mastaglia, B., Toye, C. and Kristjanson, L. J. (2003) 'Ensuring Content Validity in Instrument Development: Challenges and Innovative Approaches', *Contemporary Nurse*, 14/3: 281–91.

Molas-Gallart, J. (2015) 'Research Evaluation and the Assessment of Public Value', *Arts and Humanities in Higher Education*, 14/1: 111–26.

Molas-Gallart, J. et al. (2016) 'Towards an Alternative Framework for the Evaluation of Translational Research Initiatives', *Research Evaluation*, 25/3: 235–43.

Nunnally, J. C. and Bernstein, I. H. (1994) *Psychometric Theory*. New York: McGraw-Hill.

Ochsner, M., Hug, S. E. and Daniel, H.-D. (2012) 'Indicators for Research Quality in the Humanities: Opportunities and Limitations', *Bibliometrie - Praxis und Forschung*, 1/4: 1-17. DOI: 10.5283/bpf.157.

Osburn, H. G. (2000) 'Coefficient Alpha and Related Internal Consistency Reliability Coefficients', *Psychological Methods*, 5/3: 343–55.

Panel on the return on investments in health research. (2009) *Making an Impact: A Preferred Framework and Indicators to Measure Returns on Investment in Health Research*. Ottawa, ON (Canada): Canadian Academy of Health Science (CAHS), Ed.

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*. London: Sage Publications Ltd. http://dx.doi.org/10.4135/9781473913820

Peterson, R. A. (1994) 'A Meta-Analysis of Cronbach's Coefficient Alpha', *Journal of Consumer Research*, 21/2: 381.

Poppy, G. (2015) 'Science Must Prepare for Impact', *Nature*, 526/7571: 7.

'San Francisco Declaration on Research Assessment (DORA)'. (n.d.) <https://sfdora.org/> accessed 17 Dec 2018.

Shih, T. H. and Xitao, F. (2008) 'Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis', *Field Methods*, 20/3: 249–71.

Suriñach, J. et al. (2017) *Socio-Economic Impacts of Catalan Public Universities and Research, Development and Innovation in Catalonia*. Barcelona: Catalan Association of Public Universities (ACUP).

Trostle, J. (1992) 'Research Capacity Building in International Health: Definitions, Evaluations and Strategies for Success', *Social Science & Medicine*, 35/11: 1321–4.

Universitat Oberta de Catalunya (2017) *Strategic Plan Stage II 2017-2020*. Barcelona: UOC.

Van den Broucke, S., Dargent, G. and Pletschette, M. (2012) 'Development and Assessment of Criteria to Select Projects for Funding in the EU Health Programme', *The European Journal of Public Health*, 22/4: 598–601.

Weiss, A. P. (2007) 'Reviews and Overviews Measuring the Impact of Medical Research: Moving from Outputs to Outcomes', *Psychiatry: Interpersonal and Biological Processes*, 164/February: 206–14.

Weißhuhn, P., Helming, K. and Ferretti, J. (2018) 'Research Impact Assessment in Agriculture—A Review of Approaches and Impact Areas', *Research Evaluation*, 27/1: 36–42.

Wooding, S. et al. (2010) *Mapping the Impact: Exploring the Payback of Arthritis Research*. Santa Monica, CA: RAND Corporation.

Wooding, S. et al. (2014) 'Understanding Factors Associated with the Translation of Cardiovascular Research: A Multinational Case Study Approach', *Implementation Science*, 9/1: 47.

ANNEX 3. DEFINIONS DE LES VALORACIONS QUALITATIVES

Les definicions s'han mantingut en el llenguatge de l'avaluació (anglés).

Fase 1. Projecte anonimitzat

Finançable: Recommended indicates an outstanding proposal and translates into a very high priority for funding (first group of priority). The proposal is original, very well designed, technically feasible, practical and with a realistic work schedule. The hypothesis and goals are clearly stated. The methods section is clear, explicit, and comprehensive and data analysis is logical and well described. If a project has too many weaknesses, never could be a Recommended project.

Finançable amb reserves: Recommended with reservations indicates less enthusiasm for funding of the project. Although the proposal is feasible and meets all or most of the expected criteria (second group of priority), there are, however, a few theoretical and methodological weaknesses in it and essential information is lacking such as vaguely description of the goals, no clear hypothesis, lack of description for statistical analysis or quantification of results, etc, or a shortness of expertise of the research team. Overall, the description of the proposal is good, but somewhat confused and the research team has a strong research background and capacity in the research field. On the other hand, the proposal is very well designed, technically feasible, workable and with a realistic work schedule, but it is impossible to guarantee that all of the necessary expertise (as per the publications) is represented by the team. The proposal might be funded if appropriate recommendations are made

Dubtós: Questionable indicates a low level of scientific quality and/or relevance; therefore, it would be difficult to fund the proposal. Enthusiasm on the proposal is modest. The major weakness is the poorly developed and unfocused research design. Methodology and description of the design are questionable and superficial due to lack of details on critical elements, it is highly doubtful that the proposal will have definitive outcomes, although the research team is strong enough scientifically. On the other hand, the proposal may be not feasible, too ambitious and comprise more work than is likely to be carried out in the time frame proposed, or even the track record of the research team may be insufficient.

No finançable: Not recommended indicates significant weaknesses or absence in meeting the expected criteria; hence it would be highly difficult to fund the proposal. The proposal is not

novel, too superficial in its present form not being able to describe the exact methodology behind the study. It is also unclear what the expected outcome measures, are and there is no strong logical order of tasks and experiments. The proposal may be very ambitious and not feasible in the short time exposed. The expertise of the team may not be sufficient to undertake the project, the publication record of the PI is poor and in most cases appears irrelevant to the proposed area of research. Not recommended also indicates that the proposal does not fit with the call. (Please see the first question of the proposal evaluation form, if you answered "disagree" or "strongly disagree")

Fase 2. Projecte no anonimitzat

Finançable: Recommended indicates an outstanding proposal and translates into a very high priority for funding (first group of priority). The proposal is original, very well designed, technically feasible, practical and with a realistic work schedule. The hypothesis and goals are clearly stated. The methods section is clear, explicit, and comprehensive and data analysis is logical and well described. If a project has too many weaknesses, never could be a Recommended project.

Finançable amb reserves: Recommended with reservations indicates less enthusiasm for funding of the project. Although the proposal is feasible and meets all or most of the expected criteria (second group of priority), there are, however, a few theoretical and methodological weaknesses in it and essential information is lacking such as vaguely description of the goals, no clear hypothesis, lack of description for statistical analysis or quantification of results, etc, or a shortness of expertise of the research team. Overall, the description of the proposal is good, but somewhat confused and the research team has a strong research background and capacity in the research field. On the other hand, the proposal is very well designed, technically feasible, workable and with a realistic work schedule, but it is impossible to guarantee that all of the necessary expertise (as per the publications) is represented by the team. The proposal might be funded if appropriate recommendations are made

Dubtós: Questionable indicates a low level of scientific quality and/or relevance; therefore, it would be difficult to fund the proposal. Enthusiasm on the proposal is modest. The major weakness is the poorly developed and unfocused research design. Methodology and description of the design are questionable and superficial due to lack of details on critical elements, it is highly doubtful that the proposal will have definitive outcomes, although the research team is

strong enough scientifically. On the other hand, the proposal may be not feasible, too ambitious and comprise more work than is likely to be carried out in the time frame proposed, or even the track record of the research team may be insufficient.

No finançable: Not recommended indicates significant weaknesses or absence in meeting the expected criteria; hence it would be highly difficult to fund the proposal. The proposal is not novel, too superficial in its present form not being able to describe the exact methodology behind the study. It is also unclear what the expected outcome measures, are and there is no strong logical order of tasks and experiments. The proposal may be very ambitious and not feasible in the short time exposed. The expertise of the team may not be sufficient to undertake the project, the publication record of the PI is poor and in most cases appears irrelevant to the proposed area of research. Not recommended also indicates that the proposal does not fit with the call. (Please see the first question of the proposal evaluation form, if you answered “disagree” or “strongly disagree”)